**DSBDA Endsem – 2024, 23, 22 PYQ and IMP Q&A**

**UNIT - 3**

**Q.1 What is the data Preparation phase in Data Analytics Lifecycle. What is the Analytics Sandbox and ETLT process in this phase? [8]**

**Q2. List out different stakeholders of an analytics project. What they usually expect at the conclusion (key outputs) of a project? [8]**

**Q3. List out the activities to be carried out in model planning and model building phase. What are different tools used for these phases? [8]**

**Q4. What is linear regression, and what are its primary objectives? What is the difference between simple linear regression and multiple linear regression? How do you evaluate the performance of linear regression? [8]**

**Q5. Explain Data Analytics Cycle with suitable diagram and its phases. [8]**

**Q6. List and Explain the various activities involved in identifying potential data resources as a part of discovery phase in Data Analytics Life Cycle? [9]**

**Q7. List and explain the key roles for successful analytics project. [8]**

**Q8. Write short note on : i) Common Tools for the Model Building ii) Model selection for Data Analytics. [9]**

**Q9. What is Model Building elaborate this phase of data analytics with the help of suitable example. [9]**

**Q10. Explain any three sources of Big Data. Differentiate BI versus Data science. [8]**

**Q11. What are the three characteristic of Big Data and what are the main consideration in processing Big Data. [8]**

**Q12. Explain Descriptive, Diagnostic, Predictive analytics. [9]**

**Q13. Draw the diagram of data analytics life cycle in big data and briefly explain its phases. [8]**

**Q14. Explain in detail how the model building phase is built by team in data analytics life cycle? [9]**

**Q15. List and explain the steps in data preparation phase of data analytics life cycle. [8]**

**Q16. Write short note on the following: i) ETL ii) Common tools for the model building. iii) Model selection for data analytics. [9]**

**Q17. What is driving data deluge? Explain with one example. [9]**

**Q18. What is data science? Differentiate between Business Intelligence and Data Science. [9]**

**Q19. What are the sources of Big Data. Explain model building phase with example. [9]**

**Q20. Explain big data analytics architecture with diagram. What is data discovery phase. Explain with example. [9]**

**UNIT- 4**

**Q1. What is logistic regression, and how does it differ from linear regression? What is the sigmoid function, and what role does it play in logistic regression? [9]**

**Q2. Suppose you are given a dataset containing information about whether emails are spam or not spam, along with two features: the presence of the word "offer" (1 for present, 0 for absent) and the presence of the word "free" (1 for present, 0 for absent). You are tasked with classifying a new email with the following feature values: "offer"=1 and "free"=1. [9]**

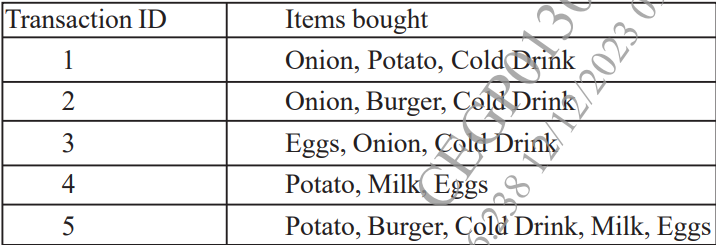
**Given the training dataset:**

**Calculate the probability that the new email is spam using Naive Bayes.**

**Q3. How does the Apriori algorithm discover frequent itemsets in a dataset? What is the role of support and confidence in the context of association rule mining using the Apriori algorithm? [9]**

**Q4. Explain the process of building a decision tree? What are the criteria used for splitting nodes in a decision tree? [9]**

**Q5. List and explain the various types of analytics in Big data. [9]**

****

**Q6. Calculates the support and confidence value for all the possible item sets.[9]**

**Q7. Explain the need of logistic regression along with its various types. [9]**

**Q8. Explain the following terms with suitable example. i) Removing Duplicates from dataset. ii) Handling Missing Data [9]**

**Q9. Explain why decision tree are used. Draw a sample decision tree and explain its parts. [9]**

**Q10. How Apriori Algorithm works, explain with suitable example? [9]**

**Q11. What is data preprocessing? Explain in details about handling missing data and transformation of data. [9]**

**Q12. Explain Naïve Bayes’ classifier and it applications. [9]**

**Q13. What are the types of analytics in big data? Explain in brief. [9]**

**Q14. Write short note on the following: i) Removing duplicates from data set. ii) Handling missing data iii) Data transformation. [9]**

**Q15. Explain various data pre-processing steps. Discuss essential python libraries for preprocessing. [8]**

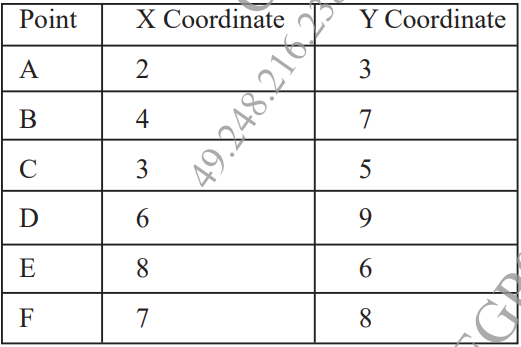
**Q16. What are association rules? Explain Apriori Algorithm in brief. [9]**

**Q17. Explain the following i) Linear Regression ii) Logistic Regression [8]**

**Q18. Explain scikit-learn library for matplotlib with example. [9]**

**UNIT – 5**

**Q1. Suppose you have the following dataset containing the coordinates of points in a 2-dimensional space: [9]**



**Perform K-means clustering on this dataset with K = 2. Assume the initial centroids to be (2,3) and (8,6). Compute the new centroids after each iteration until convergence, and assign points to their nearest centroids.**

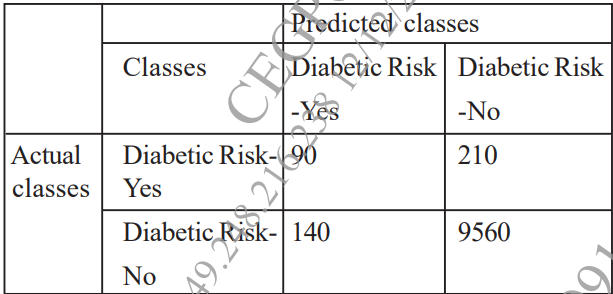
**Q2. How do you handle noise and irrelevant information in text data during preprocessing? Explain the terms bag of words and TF IDF in text analytics. [9]**

**Q3. Explain how hierarchical clustering can be used for visualizing hierarchical relationships in data with suitable example? What are some real-world applications of hierarchical clustering? [9]**

**Q4. What is the holdout method, and how does it work? Explain the difference between training set, validation set, and test set in the holdout method.[9]**

**Q5. Suppose that the given data the task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1 (2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. [8] Use the k-means algorithm to show only show only the first round of execution with cluster center.**

**Q6.Explain the following Text Analysis steps with suitable example [9] i) Part-of-speech(POS)tagging ii) Lemmatization   
Q7. Given the confusion matrix, Calculate Accuracy, Precision, Recall, Error rate with description on Diabetic Risk. [8]**



**Q8. Explain the Text Preprocessing steps with suitable example. [9]**

**Q9. What is text processing? Explain TF-IDF with example. [8]**

**Q10 .With suitable example ,explain the steps involved in k-means algorithm. [9] Q11. Define following terms with respect to confusion matrix : [8] i) Accuracy ii) Precision iii) Recall iv) AUC-ROC**

**Q12. Explain k-fold Cross Validation & Random Subsampling. [9]**

**Q13. Suppose that the given data the taste is to cluster points (With (x.y) representing location) into three cluster, where the points are. A1(2,10), A2(2,5), A3(8,4), B1 (5,8) B2(7,5) B3(6,4), C1(1,2), C2(4,9) The distance function is Euclidean distance suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. use the K-means algorithm to show only the three cluster centers after the first round of execution with steps. [9]**

**Q14. Explain the following text analysis steps with suitable example. [8] i) Part of speech (POS) tagging ii) Lemmatization iii) Stemming**

**Q15. Write short note on i) Time series Analysis ii) TF - IDF. [9]**

**Q16. What is clustering? With suitable example explain the steps involved in k - means algorithm. [9]**

**Q17. Write short note on i) Confusion matrix ii) AVC - ROC curve [9]**

**Q18. Discuss Holdout method and Random Sub Sampling methods. [9]**

**UNIT – 6**

**Q1. What is a histogram? How is it used to visualize the distribution of data? How is it different from a density plot? [9]**

**Q2. What is the Hadoop ecosystem, and what are its primary components? What is MapReduce, and how does it fit into the Hadoop ecosystem?[9]**

**Q3. What is a box plot? Explain the different components of a box plot? How do you interpret the median, quartiles, and whiskers in a box plot? What does the interquartile range (IQR) represent in a box plot? [9]**

**Q4. Explain the role of Apache Pig in data processing workflows on Hadoop? What is Apache Spark, and how does it complement Hadoop for big data processing? [9]**

**Q5. List the few data visualization tools and discuss any four applications of data visualization along with the use of the various plots with Python/R or suitable tool. [9] Q6. List the challenges of Data Visualization. Explain the types of visualization with example. [9]**

**Q7. Explain in detail the Hadoop Ecosystem with suitable diagram along with the various components. [9]**

**Q8. Write a short note on the following. [9] a) Map Reduce b) Pig**

**Q9. With a suitable example, draw a Histogram, boxplot and explain its usages. [9] Q10. Describe the data visualization tool Tableau. List of data visualization tools. [9] Q11. What is Data Visualization? Describe the challenges of data visualization. [9] Q12. Explain architecture of Apache-Pig. [9]**

**Q13. List the data visualization tools and discuss any four applications of data visualization along with the use of the suitable plot. [9]**

**Q14. List the challenges of data visualization explain the types of visualization with example. [9]**

**Q15. Explain in detail the Hadoop Ecosystem with suitable diagram [9]**

**Q16. Write a short note on the following [9] i) Map reduce. ii) Pig iii) Hive**

**Q17. With a suitable example explain Histogram and explain its usages. [8]**

**Q18. Describe the Data visualization tool “Tableau”. Explain its applications in brief. [9]**

**Q19. With a suitable example explain and draw a Box plot and explain its usages. [8] Q20. Describe the challenges of data visualization. Draw box plot and explain its usages. [9]**