**Subject-Data Science and Big Data Analytics**

**Question Bank**

| 1. Define the following terms:   Apriori property  Support  Confidence  residual sum of square  regression sum of square   1. 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.  Use the k-means algorithm to show only   * 1. The three cluster centers after the first round of execution.   2. The final clusters  1. Explain the following Text Analysis steps with suitable example    * 1. Part-of-speech(POS)tagging      2. Lemmatization      3. Stemming 2. Given the confusion matrix, Calculate Accuracy, Precision, Recall, Error rate with description on Cancer Risk.    * 1. Predicted classes    1. Classes Cancer Risk -Yes Cancer Risk  -No   Actual classes Cancer Risk-Yes 90 210   * 1. Cancer Risk-No 140 9560  1. Explain the following Text Analysis terms with suitable example.    * 1. Porter’s Stemming Algorithm      2. TFIDF 2. 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. 3. List the challenges of Data Visualization. Explain the types of visualization with example. 4. Draw the diagram of data analytics life cycle in big data and briefly explain its phases. 5. Explain in Detail how Model Building phase is built by team in Data Analytics Life Cycle? 6. What are different types of data analytics methods? List different steps for data preprocessing. 7. What is the difference between data analysis and data analytics? Explain the importance and need of data preprocessing? 8. Consider the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}    * 1. Find the least square regression line for the given data points.      2. Plot the given points and the regression line in the same rectangular system of axes. 9. Explain any four essential python libraries for big data analysis with suitable example. 10. What is Elbow Plot, explain its significance 11. Write a note on (i) Confusion Matrix (ii) AUC-ROC Curve 12. Considering a dataset and a model create confusion matrix for model evaluation algorithm 13. b) Write a note on (i) Social Network Analysis (ii) Business Analysis 14. a) What are the various tools used in Data Visualization 15. b) Illustrate Data Visualization of Scatter plot and Box Plot using Python 16. Illustrate Data Visualization of Line plot and Histogram using Python 17. Write a note on (i) Map Reduce (ii) Hive 18. Explain in detail the Hadoop Ecosystem with suitable diagram along with the various components. 19. Write a short note on the following.   Map Reduce  Pig  Hive |
| --- |
| 1. Explain the Major steps of decision tree classification |
| 1. Explain the need of logistic regression. Define Logit function and its relevance to sigmoid function. |
| 1. How does POS help in Document preprocessing? Give an example for the same. |
| 1. Explain the Porter Algorithm  used for stemming. |
| 1. What are the different methods of clustering? Give the steps of k-means clustering. |
| 1. List the few data visualization tools and discuss any two applications of data   visualization.   1. Which Python package should you use for data visualization?  Explain one package with example 2. What is Heat Map. What is the purpose of using Heat Maps. Give an example |
| 1. List  and discuss the types of data visualization. |