|  |  |  |
| --- | --- | --- |
| **Question Bank** | | |
| **BID-703- Advanced Data Science** | | |
|  | **Unit : \_\_\_\_I\_\_\_\_\_\_\_\_** | **Marks** |
|  | What is the difference between np.sort() and np.argsort()? | 2 |
|  | Explain the concept of broadcasting in NumPy in one sentence. | 2 |
|  | What is manifold learning, and how does it differ from traditional dimensionality reduction techniques? | 2 |
|  | Outline approximation algorithms. | 2 |
|  | What is hyperparameter tuning, and why is it necessary in machine learning models? | 2 |
|  | Discuss feature engineering and explain its significance in machine learning. | 2 |
|  | What is confusion matrix ? Explain. | 2 |
|  | Describe the role of train\_test\_split in model validation using Scikit-learn. | 2 |
|  | What are the three matrices produced by the Singular Value Decomposition of a matrix? | 3 |
|  | What are aggregation functions in NumPy? Explain with examples of sum(), mean(), and min() functions applied on a 2D array. | 3 |
|  | Explain the Central Limit Theorem, and why is it important in statistics. | 3 |
|  | Name two advantages of using dimensionality reduction in machine learning models. | 3 |
|  | Discuss the three matrices produced by the Singular Value Decomposition of a matrix. | 3 |
|  | Explain the null hypothesis in an A/B testing experiment. | 3 |
|  | Explain the structure of a decision tree and how it makes decisions based on input features. | 4 |
|  | Describe the process of spectral clustering and its advantages over k-means clustering. | 4 |
|  | Describe the k-means clustering algorithm in brief. | 4 |
|  | What is linear regression, and how is it used in predictive modeling? | 4 |
|  | Explain the structure of a Random Forest tree and how it makes decisions based on input features. | 4 |
|  | What is hyperparameter tuning, and why is it necessary in machine learning models? | 4 |
|  | Discuss the best parameters from the grid search, evaluate the model's performance on test data, and explain the importance of hyperparameter tuning in improving model accuracy with examples. | 6 |
|  | Discuss how spectral clustering outperforms k-means for non-linear clusters. Support your answer with examples. | 6 |
|  | Explain what is exploratory Data Analysis and its role in data science applications. | 6 |
|  | What is a recommendation system? Briefly describe how it can benefit e-commerce websites. | 6 |
|  | Compare bagging and boosting in terms of how they handle model variance and bias. | 6 |
|  | Discuss collaborative filtering and content-based filtering in recommendation systems. Provide examples of how each works. | 6 |
|  | **Unit : \_\_\_\_II\_\_\_\_\_\_\_\_** | **Marks** |
|  | What is a Jupyter Notebook, and how is it helpful in data science? | 2 |
|  | Why is Python a preferred language for data science? | 2 |
|  | What are structured arrays in NumPy? | 2 |
|  | Explain how broadcasting works in NumPy ? | 2 |
|  | Explain a DataFrame in Pandas? How does it differ from a Series? Give examples | 2 |
|  | Examine the function is used to display a plot in Matplotlib. Give example. | 2 |
|  | Write a Python function that takes a variable and returns its data type (e.g., int, float, str, list, etc.). | 2 |
|  | Write a Python function that calculates and plots the Cumulative Distribution Function (CDF) of a given NumPy array of data. | 2 |
|  |  | 3 |
|  | Which Python library would you use for building machine-learning models? Name two popular machine learning algorithms implemented in Python. | 3 |
|  | Explain the universal functions in NumPy. Provide an example of applying universal functions on a NumPy array. | 3 |
|  | Explain hierarchical indexing in Pandas. What advantage does it offer when working with multi-dimensional data? | 3 |
|  | Explain the purpose of the reshape() function in NumPy. Provide an example of reshaping a 1D array into a 2D array | 3 |
|  | Describe broadcasting in NumPy with an example where an operation between two arrays of different shapes is performed. | 3 |
|  | Explain the purpose of the reshape() function in NumPy. Provide an example of reshaping a 1D array into a 2D array. | 4 |
|  | Discuss how Pandas can be used to perform data manipulation operations such as filtering, grouping, and applying functions to DataFrames. Provide examples to illustrate these operations. | 4 |
|  | Illustrate a code snippet to demonstrate the use of np.sum() to calculate the sum of all elements in a 2D NumPy array. | 4 |
|  | Create a code example that demonstrates how to group a Pandas DataFrame by multiple columns and calculate both the mean and standard deviation of another column. | 4 |
|  | Write a function that sorts a NumPy array based on a custom key (e.g., absolute values) using a user-defined comparison function. | 4 |
|  | What is a time-series dataset with an example? | 4 |
|  | Write a function that merges two DataFrames based on multiple keys and resolves any conflicts in overlapping column names. Give example. | 6 |
|  | Implement a custom aggregation function that computes the weighted average of a DataFrame column based on another column with a proper example. | 6 |
|  | Create a code example that demonstrates how to decompose a time series into its trend, seasonality, and residual components using a sample dataset. | 6 |
|  | Implement a function that takes a Pandas DataFrame, identifies columns with more than 20% missing values, and drops them. Explain the importance of handling missing data. | 6 |
|  | Explain how they are useful for handling heterogeneous data and provide a code example of creating and accessing structured arrays. | 6 |
|  | Explain hierarchical indexing in Pandas. What advantage does it offer when working with multi-dimensional data? | 6 |
|  | **Unit : \_\_\_\_\_III\_\_\_\_\_\_\_** | **Marks** |
|  | Name two different operation modes of Hadoop. | 2 |
|  | What is the purpose of the mapper and reducer functions in Hadoop’s MapReduce framework? | 2 |
|  | Write a simple command to list the contents of a directory in HDFS. | 2 |
|  | What is the role of NameNode in the HDFS architecture? | 2 |
|  | Explain the role of Hadoop YARN in the Hadoop ecosystem. | 2 |
|  | Outline the steps involved in setting up a multi-node Hadoop cluster | 2 |
|  | What is Hadoop, and why is it widely used for big data processing? | 2 |
|  | What is YARN in Hadoop, and what does it stand for? | 2 |
|  | What is YARN in Hadoop, and what does it stand for? | 3 |
|  | What is Hadoop, and what is its primary use in data processing? | 3 |
|  | Describe the Hadoop, and what is its primary use in data processing? | 3 |
|  | List and briefly describe the four main modules of the Hadoop framework. | 3 |
|  | Write a Hadoop command to delete a file from HDFS and explain its syntax. | 4 |
|  | What is the purpose of a multi-node cluster in Hadoop, and how does it enhance performance? | 4 |
|  | Define the terms mapper and reducer in the context of the MapReduce programming model. | 4 |
|  | Describe the four main modules of the Hadoop framework ? | 4 |
|  | What are the different operation modes in Hadoop? Explain the differences between them. | 4 |
|  | List and briefly describe the four main modules of the Hadoop framework | 4 |
|  | What is the purpose of the mapper and reducer functions in Hadoop’s MapReduce framework? | 4 |
|  | Name two different operation modes of Hadoop. | 4 |
|  | Describe how YARN handles resource management and job scheduling in Hadoop. Include the roles of ResourceManager and NodeManager. | 6 |
|  | Briefly explain the client-server architecture of Hadoop. | 6 |
|  | Explain how Hadoop Streaming works. What are the roles of the mapper and reducer in streaming mode? | 6 |
|  | Discuss the architecture of HDFS in detail, focusing on how it achieves fault tolerance and high throughput. | 6 |
|  | Explain the role of Hadoop YARN to improve resource management. | 6 |
|  | Explain how Hadoop Streaming works. | 6 |
|  | Write and explain the syntax for the following Hadoop commands:  a) Creating a directory in HDFS  b) Copying a file from the local file system to HDFS  c) Checking the status of a file in HDFS. | 6 |
|  | Explain the concept of Hadoop Streaming and how the mapper and reducer phases can be implemented using Python or any other programming language. | 6 |
|  | **Unit : \_\_\_\_\_IV\_\_\_\_\_\_\_** | **Marks** |
|  | Describe a machine learning case study where classification algorithms are used. Mention the problem and the steps taken to solve it. | 2 |
|  | What is logistic regression, and how can it be applied to the Titanic dataset to predict survival? | 2 |
|  | What role does exploratory data analysis (EDA) play in a data science case study? | 2 |
|  | Name two applications of machine learning in healthcare. | 2 |
|  | How does YARN improve resource management, and what are its key components? | 2 |
|  | Describe how to submit a MapReduce job in Hadoop using the command line. | 2 |
|  | What is a data science case study, and why is it useful in understanding real-world problems? | 2 |
|  | Explain cross-validation and its importance in evaluating machine learning models. | 2 |
|  | What are the common challenges faced when working on a data science case study? | 3 |
|  | List three popular Python libraries used in data science. Briefly describe the purpose of each. | 3 |
|  | What are the key advantages of using online notebooks (e.g., Jupyter or Google Colab) for data science projects? | 3 |
|  | Explain the algorithms typically used and how the model can be trained on an imbalanced dataset. | 3 |
|  | Apply feature engineering to the Titanic dataset by creating new features (e.g., family size, title extraction from the name) | 3 |
|  | Apply feature engineering to the Titanic dataset by creating new features (e.g., family size, title extraction from the name) | 3 |
|  | Describe the entire pipeline, from data collection, feature selection, model building, and evaluation. | 4 |
|  | Write a Python script that applies k-nearest neighbors (KNN) to the Titanic dataset and evaluates the model performance using a confusion matrix and accuracy score. | 4 |
|  | Name two Python libraries commonly used for data manipulation and visualization. | 4 |
|  | Explain the importance of feature scaling in machine learning models, with examples of algorithms that are sensitive to feature scaling. | 4 |
|  | Compose an example of a basic workflow in a machine learning project, from data preparation to model evaluation. | 4 |
|  | Discuss one application and the machine learning algorithm commonly used for it. | 4 |
|  | How can you visualize the correlation matrix of a dataset in Python using Pandas and Matplotlib? Provide a simple code snippet. | 4 |
|  | Discuss various evaluation metrics used in classification problems (e.g., accuracy, precision, recall, F1-score). Implement and explain these metrics for a model trained on the Titanic dataset. | 4 |
|  | Implement a time series forecasting model in Python using a simple dataset (e.g temperature data). Explain the techniques used and evaluate the performance of the model. | 6 |
|  | Describe the entire pipeline, from data collection, feature selection, model building, and evaluation. | 6 |
|  | Explain how Python’s sci-kit-learn library is used to implement a decision tree classifier. | 6 |
|  | Explain the role of Hadoop YARN in the Hadoop ecosystem | 6 |
|  | Describe a machine learning case study where classification algorithms are used. Mention the problem and the steps taken to solve it. | 6 |
|  | Explain cross-validation and its importance in evaluating machine learning models. | 6 |
|  | Implement a time series forecasting model in Python using a simple dataset (e.g., stock prices ). Explain the techniques used and evaluate the performance of the model. | 6 |
|  | Write a Python script that applies k-nearest neighbour’s (KNN) to the Titanic dataset and evaluates the model performance using a confusion matrix and accuracy score. | 6 |