

**BDS 1101-BSD 3101: Principles of Data Science**

1. **a) Explain why effective communication is critical for a data scientist or analyst. Support your answer with a valid example. (2 Marks)**

It facilitates Understanding Stakeholder Needs. Helps comprehend and address specific requirements of stakeholders from different domains, ensuring accurate project alignment.

*A data scientist analyzing customer churn data communicates findings effectively, conveying not just statistical analysis but also the business implications. This ensures decision-makers understand the revenue impact and take actions, like investing in personalized recommendation features to reduce churn and increase customer retention*.

2. **b) Discuss any three effective communication techniques for data analysts.**
3. **(3 Marks)**

* Data Visualization. Use charts, graphs, and other visual representations to simplify complex data sets. Visualization enhances understanding, making it easier for non-technical stakeholders to grasp insights quickly.
* Storytelling with Data. Frame data insights within a narrative to create a compelling story. This approach helps engage the audience, making the data more relatable and memorable.
* Tailoring Communication to the Audience. Adapt communication style and level of technical detail based on the audience's background and expertise. Tailoring ensures that information is accessible and relevant to diverse stakeholders.

1. **c) Describe four most commonly used algorithm by data scientist. (4 Marks)**

* Linear Regression: Used for predicting a continuous outcome based on one or more predictor variables. It establishes a linear relationship between the input features and the target variable.
* Decision Trees: Tree-like structures where each node represents a decision based on input features. It recursively splits the data into subsets, making it a versatile algorithm for classification and regression tasks.

1. K-Means Clustering: An unsupervised learning algorithm that partitions data into k clusters based on similarity. It assigns data points to clusters with the goal of minimizing the within-cluster variation.
3. **d) Discuss the six steps in the data preparation process. (6 Marks)**
4. 1. Data Collection: Gather relevant data from various sources, ensuring it aligns with the objectives of the analysis or modeling task. The quality and relevance of collected data directly impact the success of subsequent steps.
5. 2. Data Cleaning: Identify and handle missing values, outliers, and errors in the dataset to ensure data integrity and accuracy. Clean data prevents biased analyses and improves the performance of machine learning models.
6. 3. Exploratory Data Analysis (EDA): Analyze and visualize the data to understand patterns, relationships, and potential insights. EDA guides feature selection, model choice, and helps identify key trends in the data.
7. 4. Feature Engineering: Transform and create new features to enhance the performance of machine learning models. This may include scaling, encoding categorical variables, or creating interaction terms. Well-engineered features can significantly improve a model's ability to capture patterns and make accurate predictions**.**
8. 5. Data Splitting: Divide the dataset into training and testing sets. The training set is used to train the model, while the testing set evaluates its performance on unseen data. Assessing model generalization on new data helps avoid overfitting and provides a realistic measure of predictive accuracy.
9. 6. Data Scaling and Normalization: Standardize or normalize numerical features to a common scale, ensuring that no single feature disproportionately influences the model. Scaling enhances the stability and convergence of certain algorithms, promoting better model performance**.**