**Part 1: Data Preprocessing and Preparation**

**1.(a)** The quality of model depends upon the quality of data inputted inside it, for both training data and testing data. If the data is not good enough, the model will also give poor results, and vice-versa.

**(b)** Real world data is often incomplete, messy and inconsistent. A model requires clean and understandable data. It can create complications in the output. It has to be transformed into a usable and efficient format. There is a presence of outliers, missing values and categorical data which needs to be organized before analyzing the data.

**2.** Imputation can introduce biasness if not represented accurately. It is done when there are a lot of missing values. Deletion is done when the number of missing values in data is relatively low. However, it can lead to loss of required information which might be essential for the model.

**3.** Min-Max scaling for keeping the data in the range of zero to 1. Standardization involving the mean to be zero and standard deviation to be one to ensure that all features contribute equally in a dataset, and to improve model performance. Sensitive-KMeans Clustering. Not Sensitive-Decision Trees.

**Part 2: Data Preprocessing and Preparation**

**4.**To evaluate how well our model will perform on new, unseen data. Training- “Teaching” or “Fitting” the model. Testing-Getting an unbiased estimate of model’s performance on unseen data. Validating-Optimizing the model during training process to avoid overfitting from the original data.

**5.**It refers to a situation when a model learns the test data too well, leading to it making inaccurate predictions during the new, untested data.

**(a)**Memorizes training sets and performs poorly on new data.

**(b)**To avoid any misunderstandings or inaccuracies in testing the datasets when provided to the model.

**6.**A model learns to map inputs and outputs based on labelled training data. To enhance its working and efficiency when given new data sets, as it might just memorize the datasets rendering to poor performances for other data.

**7.**It is the technique of creating new features with more information derived from ones which already exist. It can be used to create age by just inputting the date of birth, or combining height and month to calculate BMI.

**Part 3: Model Validation Techniques**

**8.**Performance can be sensitive to how the validation split was made, especially with smaller datasets. This can affect the performance of the models.

**9.(a)**It shuffles the original training data and splits it into K equal-sized folds. For each fold, it is taken as a validation set and the others are taken in the form of a testing set. It ensures more robust estimate of performance, uses data more efficiently.

**(b)**A model is trained five times.

**10.**It is evaluation of final, trained model on a completely independent dataset not used at any stage of model development. It is the purest test of generalization of model to real-world beyond its specific characteristics.

**11.**Information from the test or validation set influences the model’s training process, leading to generalized patterns and poor cases and results. The data is seen by the model which should not have happened. It might lead to overfitting if not avoided on time, leading to poor scenarios.

**Part 4: Model Deployment Concepts**

**12.**To integrate the created model to a productive environment where it can be used and accessed.

**13.**It allows for model reuse, ease of deployment and scalability without retraining the models.

**14.**Batch Predictions for Personalized Campaign Recommendations. Real-Time Predictions for Adaptive Advertising.

**15.** A developer or data scientist has written some code working fine on a local machine but breaks in a pipeline or in production. Docker packages applications and their dependencies into isolated containers.