Part 1: Data Preprocessing & Preparation

a) GIGO means Garbage In Garbage Out. If the data that goes into the model is full of mistakes, missing stuff, or just wrong, the model will learn bad patterns and give wrong answers. Even if the model is good, it can't do anything if the input is garbage.

b)

- Missing values – sometimes data is not complete. Like someone didn't fill in their age or income. Models can't use empty boxes so it needs fixing.

- Outliers – these are values way too high or low compared to others. They confuse the model and mess with accuracy.

- Categorical data – things like "red", "blue", "male", "female". Models can't understand words unless we turn them into numbers. If not done right, model might read it wrong.

2.

- Imputation means filling missing values with the average or something like that. This works when only a few values are missing. But the risk is, you are adding fake numbers.

- Deletion means removing the row or column with missing data. It's simple and safe if only a few are missing. But you might lose good data.

3.

Feature scaling makes all features more equal by putting them in similar range. Some models use distances, and if one value is too big it controls everything.

- Model that needs it: KNN

- Model that doesn't care: Decision Trees

Part 2: Model Training, Testing, and Overfitting

Splitting into train, validation, and test sets helps build better models.

- Train set is used to make the model learn.

- Validation set helps pick the best model by testing while training.

- Test set checks how good the model really is on new data. You only use it once, at the end.

5.

a) When a model overfits, it becomes too good at training data and memorizes everything. But when you give new data, it doesn't work well because it never learned general rules

b) A separate test set helps catch this. It's never seen before, so if the model does badly here, we know it only memorized the training set.

6.

Loss function tells how far off the model's guess is. The model tries to make the loss smaller so it can guess better. Training keeps changing the model to reduce this number.

7.

Feature engineering means making new features from old ones to help the model.
Example – if we have date of birth, we can make a new feature called age. Age is more useful than just the birth date.

Part 3: Model Validation Techniques

8.

One hold-out set only tests the model one way. If the split is bad or lucky, results won't be accurate. It doesn't show the full picture of how the model performs.

9.

a) In K-Fold Cross Validation, the data is split into K parts. Each part gets used as a test set once. The model trains K times on different parts. This gives a better average.
b) If K = 5, model is trained 5 times.

10.

External validation means testing on a new dataset that the model never saw before, even during training. This is better because it shows if the model really works in the real world, not just on the same kind of data.

11.

Data leakage happens when something from the test data leaks into training. For example, scaling the whole dataset before splitting. This makes results look better than they really are. It's like cheating without knowing.

Part 4: Model Deployment Concepts

Model deployment means putting the trained model into use. Like adding it to an app or website where it can make predictions for real users or systems.

13.

Saving the model lets you use it again later without training from scratch. It also helps share it with others. Tools like pickle help store the trained model with its settings.

14.

- Batch – good when you want to predict for many users at once, like every night. For example, sending emails to all users who are likely to quit.

- Real-time – when you need the prediction instantly, like when someone clicks a product and the model has to recommend something right away.

15.

“Works on my machine” means something runs fine on your computer but not on someone else’s. Docker helps by putting all the code and software into one software that works the same everywhere.