Feature Engineering & Machine Learning Assignment Answers

1. How do we handle categorical variables in Machine Learning? What are the common techniques?

Categorical variables are converted into numerical format using:

- Label Encoding
- One-Hot Encoding
- Ordinal Encoding
- Binary/Target Encoding.

2. What do you mean by training and testing a dataset?

Training involves teaching the model using known data. Testing evaluates the model's performance on new, unseen data.

3. What is sklearn.preprocessing?

Its a Scikit-learn module for preprocessing data including encoding, scaling, and normalization.

4. What is a Test set?

A reserved portion of data used only to evaluate the models performance after training.

5. How do we split data for model fitting (training and testing) in Python?

Use train_test_split() from sklearn:

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2).
```

6. How do you approach a Machine Learning problem?

Steps:

1. Define problem
2. Collect data
3. EDA
4. Preprocess
5. Train/Test Split

6. Model Selection

7. Train

8. Evaluate.

7. Why do we have to perform EDA before fitting a model to the data?

EDA helps understand data distribution, find patterns, detect outliers, and decide preprocessing techniques.

8. What is correlation?

A statistical measure of linear relationship between variables. Range: -1 to +1.

9. What does negative correlation mean?

If one variable increases, the other decreases. Example: more exercise, less weight.

10. How can you find correlation between variables in Python?

Use pandas:

df.corr()

11. What is causation? Explain difference between correlation and causation with an example.

Causation: One variable causes change in another.

Correlation: Two variables change together but may not influence each other.

Example: Ice cream sales and drowning both rise in summer (correlated, not causal).

12. What is an Optimizer? What are different types of optimizers? Explain each with an example.

Optimizers update model parameters to minimize loss.

Types: SGD, Adam, RMSProp.

Example: optimizer = tf.keras.optimizers.Adam(learning_rate=0.01)

13. What is sklearn.linear_model ?

Module for linear models like LinearRegression, LogisticRegression, Ridge, Lasso.

14. What does model.fit() do? What arguments must be given?

Trains model with data. Arguments: X_train, y_train

15. What does model.predict() do? What arguments must be given?

Predicts output using model. Argument: X_test

16. What are continuous and categorical variables?

Continuous: Numeric (e.g., height)

Categorical: Categories (e.g., gender)

17. What is feature scaling? How does it help in Machine Learning?

Standardizes feature range. Helps model convergence and accuracy.

18. How do we perform scaling in Python?

Use StandardScaler or MinMaxScaler:

from sklearn.preprocessing import StandardScaler

X_scaled = scaler.fit_transform(X)

19. What is sklearn.preprocessing?

Module for preprocessing tasks like encoding and scaling.

20. How do we split data for model fitting (training and testing) in Python?

Use train_test_split() from sklearn.

21. Explain data encoding?

Converts categories to numbers using Label Encoding, One-Hot Encoding, etc.