**INTERVIEW QUESTIONS/ANSWERS**

**1. What are the different types of missing data?**

* **Missing Completely at Random (MCAR):** The fact that data is missing has nothing to do with the data itself—there is no pattern or reason why values are absent.
* **Missing at Random (MAR):** The missingness is related to some observed data but not to the value of the missing data itself. For example, age values might be missing more often for people with higher incomes.
* **Missing Not at Random (MNAR):** The missingness is related to the value of the missing data, even after accounting for other observed data; for example, people with very low salaries may be less likely to report their salaries.

**2. How do you handle categorical variables?**  
There are several strategies, including:

* **One-Hot Encoding:** Create a new column for each category and use 0 or 1 to indicate its presence.
* **Label Encoding:** Assign each unique category a numerical value. Best used for ordinal data where order matters.
* **Frequency or Target Encoding:** Replace each category with its frequency or its relation to the target outcome.
* **Ordinal Encoding:** Used when categories have a natural order (e.g., “Low”, “Medium”, “High”).

**3. What is the difference between normalization and standardization?**

* **Normalization** scales all values in a feature to a range between 0 and 1 (or -1 to 1), preserving the shape but changing the scale; typically used when you want to bound data (e.g., using MinMaxScaler).
* **Standardization** transforms the data to have a mean of 0 and a standard deviation of 1, centering and scaling it to resemble a standard normal distribution; this is useful when algorithms assume features are centered around zero.

**4. How do you detect outliers?**  
Outliers can be found through:

* **Visualization:** Box plots and scatter plots visually reveal points lying far from the rest.
* **Statistical Methods:** Z-scores (standard deviations from mean), or using the IQR (Interquartile Range—values outside 1.5\*IQR from Q1 or Q3 are flagged).
* **Model-Based:** Algorithms like Isolation Forest or DBSCAN can also identify points that do not fit the pattern.

**5. Why is preprocessing important in ML?**  
Preprocessing transforms raw data into a clean, structured format. This is crucial because:

* Algorithms perform better on high-quality, consistent input.
* It helps handle missing values, scale differences, and converts non-numeric data for ML models.
* Proper preprocessing leads to more accurate, reliable, and interpretable results.

**6. What is one-hot encoding vs label encoding?**

* **One-Hot Encoding:** Converts categories into separate binary columns (one column per category). Removes any sense of order and is best for categorical data where no order exists.
* **Label Encoding:** Assigns an integer value to each category, making it a single numerical column. Suitable for ordinal data, but can mislead tree/linear models if data is truly nominal.

**7. How do you handle data imbalance?**  
Common approaches are:

* **Resampling:** Oversample the minority class or undersample the majority class.
* **Synthetic Data Generation:** Use techniques like SMOTE to create new synthetic examples for the minority class.
* **Algorithmic Approach:** Use models robust to imbalance or adjust class weights (e.g., in scikit-learn, setting class\_weight='balanced').
* **Evaluation Metrics:** Use metrics such as F1-score, precision-recall, or ROC-AUC instead of just accuracy.

**8. Can preprocessing affect model accuracy?**  
Absolutely—effective preprocessing leads to higher accuracy by ensuring the data is consistent, relevant, and ready for modeling. Poor or neglectful preprocessing, on the other hand, often results in less accurate, unreliable, or even biased models. The way data is cleaned, scaled, and encoded plays a huge role in how well the model learns and generalizes.