

# **DATA PREPROCESSING**

## **MISSING DATA HANDLING**

1. What is missing data? What are its types?
2. What is MCAR (Missing Completely At Random)?
3. What is MAR (Missing At Random)?
4. What is MNAR (Missing Not At Random)? Give an example.
5. How do you detect missing data? (isnull(), isna())
6. Why is missing data pattern analysis important?
7. How do you calculate missing data percentage?
8. What percentage of missing data is acceptable in a column?
9. How do you visualize missing data? (missing no library)
10. How do you create a missing data report?

## **Imputation Techniques**

11. What is mean imputation? When should you use it?
12. When is median imputation better than mean?
13. When do you use mode imputation?
14. What are the limitations of mean imputation?
15. What are forward fill and backward fill? (Time series data)
16. What is interpolation? What are its types (linear, polynomial)?
17. What is KNN imputation? How does it work?
18. What is MICE (Multiple Imputation by Chained Equations)?
19. What is regression imputation?
20. When is it useful to make missing values a separate category?

## **OUTLIER DETECTION & TREATMENT**

### **Outlier Detection Methods**

21. What is an outlier? Give an example.
22. Why are outliers harmful for models?
23. When can outliers be beneficial?
24. How do you detect outliers using visual methods? (Boxplot, scatter plot)
25. Explain the IQR (Interquartile Range) method.
26. IQR formula:  $Q3 - Q1$

27. State the steps for outlier detection using IQR method.
28. What are the lower bound and upper bound formulas in IQR method?
29. What is the Z-score method for outlier detection?
30. Explain the Z-score  $> 3$  or  $< -3$  rule.
31. What is Modified Z-score (using median)?
32. What is the Isolation Forest algorithm?
33. How do you detect outliers using DBSCAN clustering?
34. What is Local Outlier Factor (LOF)?
35. What is Cook's distance in regression?
36. What are leverage points?
37. What are multivariate outliers? How are they different from univariate?
38. What is Mahalanobis distance?

### **Outlier Treatment**

39. When is it appropriate to remove outliers?
40. What is capping/flooring of outliers? (Winsorization)
41. How does log transformation help in handling outliers?
42. Why is robust scaling better for outliers?
43. Why is domain knowledge important in outlier treatment?
44. When is it right to treat outliers as missing values?

### **FEATURE SCALING**

#### **Normalization**

45. Why is feature scaling necessary?
46. What is normalization? State the formula.
47. Min-Max scaling formula:  $(x - \min) / (\max - \min)$
48. What is the range of normalization? (0 to 1)
49. When should you use normalization?
50. Is normalization affected by outliers?

#### **Standardization**

51. What is standardization? State the formula.
52. Z-score normalization formula:  $(x - \mu) / \sigma$
53. What is the output distribution of standardization? (mean=0, std=1)

54. What is the difference between standardization and normalization?

55. When should you use standardization and when normalization?

### **Other Scaling Methods**

56. What is Robust Scaler? State the formula.

57. Why is Robust Scaler better for outliers?

58. What is MaxAbsScaler?

59. What is unit vector scaling / L2 normalization?

60. What is power transformation? (Box-Cox, Yeo-Johnson)

61. What is the limitation of Box-Cox transformation?

62. How is Yeo-Johnson transformation different from Box-Cox?

63. When do you use log transformation?

64. What is square root transformation?

65. Why should scaling not be done separately on train and test data?

## **ENCODING CATEGORICAL VARIABLES**

### **Basic Encoding Techniques**

66. What are categorical variables? What are their types (Nominal, Ordinal)?

67. What is the difference between nominal and ordinal variables?

68. Why do we need to convert categorical variables to numerical?

69. What is Label Encoding? Give an example.

70. What is the limitation of Label Encoding?

71. When should you use Label Encoding?

72. What is ordinal encoding? How is it different from label encoding?

73. What is One-Hot Encoding? Explain in detail.

74. How does One-Hot Encoding work? Give an example.

75. What is the dummy variable trap?

76. How do you avoid the dummy variable trap in One-Hot Encoding?

77. What is the high cardinality problem in One-Hot Encoding?

78. What is Binary Encoding?

79. How is Binary Encoding better than One-Hot for high cardinality?

### **Advanced Encoding Techniques**

80. What is Frequency Encoding / Count Encoding?

81. What is Target Encoding / Mean Encoding?
82. What is the problem with target encoding? (Data leakage)
83. How do you avoid overfitting in target encoding?
84. What is Leave-One-Out Encoding?
85. What is Weight of Evidence (WoE) encoding?
86. What is hash encoding / feature hashing?
87. What is embedding encoding? (Deep learning context)
88. What is BaseN encoding?
89. What is Helmert encoding?
90. What is backward difference encoding?

## **FEATURE TRANSFORMATION**

### **Distribution Transformation**

91. What is a skewed distribution? Left skew vs Right skew.
92. How do you measure skewness?
93. How does log transformation reduce skewness?
94. Natural log vs Log10 - when should you use which?
95. What is square root transformation? When is it better than log?
96. What is cube root transformation?
97. What is reciprocal transformation?
98. Explain Box-Cox transformation in detail.
99. What is the Box-Cox lambda parameter?
100. When is Yeo-Johnson transformation necessary?

### **Binning / Discretization**

101. What is binning? Why do we do it?
102. What is equal width binning?
103. What is equal frequency binning?
104. What is custom binning? Give an example.
105. Give a practical example of converting age into bins.
106. Does binning cause information loss?
107. What is quantile-based discretization?
108. What is K-means binning?

109. What is decision tree-based binning?

110. How do you decide the optimal number of bins?

## **HANDLING IMBALANCED DATA**

### **Imbalanced Data Basics**

111. What is an imbalanced dataset? What is the problem?

112. How do you measure class imbalance ratio?

113. What are majority class and minority class?

114. How does imbalanced data affect models?

115. Where is imbalanced data commonly found? (Fraud, disease detection)

### **Resampling Techniques**

116. What is under sampling? When should you use it?

117. What is random undersampling?

118. What are Tomek links?

119. What is Edited Nearest Neighbors (ENN)?

120. What is oversampling?

121. What is random oversampling? What is its problem?

122. What is SMOTE (Synthetic Minority Over-sampling Technique)?

123. How does SMOTE work? Explain in detail.

124. What is ADASYN?

125. What is Borderline-SMOTE?

### **Other Techniques**

126. What is the SMOTE + Tomek links combination?

127. What is SMOTE + ENN?

128. What is adjusting class weights?

129. What is cost-sensitive learning?

130. What are ensemble methods (BalancedRandomForest, EasyEnsemble)?

## **DATE & TIME FEATURES**

### **DateTime Extraction**

131. Why are DateTime features important?

132. What features can you extract from a date column?

133. How do you extract year, month, and day?

- 134. How do you extract day of week? (Monday = 0)
- 135. How do you extract quarter?
- 136. How do you create an is\_weekend feature?
- 137. What are is\_month\_start and is\_month\_end features?
- 136. How do you extract hour, minute, and second? (Time series)
- 137. What is week of year?
- 138. What is day of year feature?

### **Advanced DateTime Features**

- 139. What are cyclical features? (Month, Hour) Why are they necessary?
- 140. How do you perform cyclical encoding? (sin, cos transformation)
- 141. What is time since/until a specific date feature?
- 142. How do you add holiday features?
- 143 .What is the difference between business days and calendar days?
- 144. What are lag features in time series?
- 145. What are rolling window features?
- 146. What are time-based aggregations?

## **TEXT DATA PREPROCESSING**

### **Text Cleaning**

- 147. Why is text preprocessing necessary in NLP?
- 148. Why do we lowercase text?
- 149. How do you remove punctuation?
- 150. How do you remove special characters and numbers?
- 151. How do you remove extra whitespaces?
- 152. How do you remove HTML tags?
- 153. How do you remove URLs and email addresses?
- 154. How do you handle emojis?

### **Tokenization & Normalization**

- 155. What is tokenization? Word-level vs character-level.
- 156. What are stopwords? Why do we remove them?
- 157. Give examples of common stopwords (the, is, at, which).
- 158. What are domain-specific stopwords?

159. What is stemming? Which algorithm is used? (Porter Stemmer)
160. What is lemmatization?
161. What is the difference between stemming and lemmatization?
162. Why is lemmatization accurate but slow?
163. What are N-grams? Explain bigrams and trigrams.

### **Text Vectorization**

164. What is Bag of Words (BoW)?
165. What are the limitations of BoW?
166. What is TF-IDF? What is the full form?
167. What is Term Frequency (TF)?
168. What is Inverse Document Frequency (IDF)?
169. Why is TF-IDF better than BoW?
170. What are word embeddings? (Word2Vec, GloVe)
171. What is the difference between CountVectorizer and TfidfVectorizer?

### **DATA QUALITY & VALIDATION**

#### **Data Quality Checks**

172. What data quality issues can occur?
173. How do you detect duplicate rows?
174. What should you keep in mind when removing duplicate rows?
175. What is data type validation?
176. How do you automatically detect data types? (infer objects)
177. What is the problem with mixed data types in a column?
178. What are invalid values? Give an example.
179. How do you handle negative values where they shouldn't exist?
180. How do you detect out of range values?
181. What is inconsistent formatting? (dates, phone numbers)
182. How do you detect typos and spelling errors?

#### **Data Validation**

183. What is schema validation?
184. What is constraint checking? (min, max, allowed values)
185. What is cross-field validation? Give an example.

186. What is referential integrity check?

187. What is business rules validation?

188. What is data profiling?

## **ADVANCED PREPROCESSING**

### **Feature Interaction**

189. What is feature interaction?

190. What are polynomial features? Give an example.

191. How do you create interaction terms manually?

192. What is the effect of the degree parameter in Polynomial Features?

193. What are feature crosses?

### **Dimensionality Reduction (Preprocessing Context)**

194. What are the problems of high dimensionality?

195. What is the curse of dimensionality?

196. What is the difference between feature selection and feature extraction?

197. How do you identify redundant features using a correlation matrix?

198. What is the variance threshold method?