Cheatsheet: Data Pre-processing

1. Data Sources

- **Popular Datasets:** OpenML, Kaggle, UCI ML Repository, AWS, TensorFlow datasets.
- **Meta Portals:** DataPortals.org, OpenDataMonitor.eu.

2. End-to-End ML Project Steps

1. **Frame the Problem**
   - Define the problem
   - Assess if ML is the right approach
   - Identify current solutions & their limitations
2. **Data Collection & Cleaning**
   - Handle missing values: Removal, Mean/Median Imputation
   - Remove duplicate/zero variance columns
   - Detect and handle outliers
3. **Data Transformation**
   - **Feature Scaling:**
     - **Min-Max Scaling**: Rescales to [0,1]
     - **Standardization**: Zero mean, unit variance
   - **Feature Engineering:**
     - Bucket data
     - Add new features
     - Gaussian RBF transformation
4. **Data Splitting & Validation**
   - **Train/Test Split:** Typically 80/20 split
   - **Cross-validation:** k-fold, Leave-One-Out
5. **Model Training & Selection**
   - **Performance Metrics:** RMSE, MAE, R^2
   - **Hyperparameter Tuning:** GridSearchCV, RandomizedSearch
   - **Feature Importance:** Drop less relevant features
6. **Final Model Evaluation**
   - Compare with baseline methods
   - Ensure test data remains unseen
   - Deploy model

MCQs

1. Which performance metric is more sensitive to outliers?
a) Mean Absolute Error (MAE)
b) Root Mean Square Error (RMSE)
c) Median Absolute Error
d) None of the above
Answer: (b) RMSE

2. What is the main advantage of RandomizedSearchCV over GridSearchCV?
a) It guarantees finding the best hyperparameter combination
b) It randomly selects hyperparameters for faster optimization
c) It always outperforms GridSearchCV
d) It only works for continuous hyperparameters
Answer: (b)

3. Which of the following is NOT a method of handling missing values?
a) Removing the entire dataset
b) Mean/Median Imputation
c) Filling with random values
d) Using KNN imputation
Answer: (a)

4. Which function in Scikit-Learn is used for train-test splitting?
a) train_test_split()
b) split_data()
c) model_selection_split()
d) data_partition()
Answer: (a)

5. What is the correlation coefficient range in Pearson's r?
a) 0 to 1
b) -∞ to +∞
c) -1 to +1
d) None of the above
Answer: (c)

Subjective Questions

1. Explain the importance of data pre-processing in machine learning.
2. Compare and contrast Min-Max Scaling and Standardization. When should each be used?
3. How does cross-validation help in model evaluation? Explain with an example.
4. Discuss different methods for handling missing values. Which one is best suited for structured data?
5. Explain the concept of feature importance and its role in model optimization.
6. Describe the role of hyperparameter tuning in model selection. How does GridSearchCV work?
7. Discuss the significance of outlier detection. Mention two statistical methods for outlier removal.
8. Explain how Pearson's correlation coefficient is useful in feature selection. Provide an example.
9. What are the steps involved in an ML pipeline, from raw data to model evaluation?
10. Why is it essential to keep the test data separate from training data in ML?