■ Scikit-Learn Learning Checklist

1. Setup & Fundamentals

- [ ] Install scikit-learn, check version, understand dependencies (NumPy, SciPy)
- [ ] Familiarize with the API design: .fit(), .predict(), .transform(), .fit_transform()
- [ ] Understand basic ML workflow: data → preprocess → train/test → model → evaluate → tune
- [ ] Feature engineering vs model building vs evaluation
- [ ] Understand scikit-learn’s use-cases and limitations

2. Data Preparation & Feature Engineering

- [ ] Handling missing values (SimpleImputer)
- [ ] Encoding categorical variables (OneHotEncoder, OrdinalEncoder)
- [ ] Scaling/normalization (StandardScaler, MinMaxScaler)
- [ ] Feature transformation (log, power, Box-Cox)
- [ ] Dimensionality reduction (PCA, TruncatedSVD)
- [ ] Feature selection (SelectKBest, RFE)
- [ ] Train/test split (train_test_split, stratify)
- [ ] Pipelines & ColumnTransformer
- [ ] Avoiding data leakage

3. Supervised Learning: Regression & Classification

- [ ] Regression: Linear, Ridge, Lasso, SVR, DecisionTree, RandomForest, GradientBoosting
- [ ] Classification: LogisticRegression, KNN, SVC, DecisionTree, RandomForest, Naive Bayes, AdaBoost
- [ ] Handling imbalanced classes, class weights
- [ ] Model interpretability: coefficients, feature importance, PDP plots

4. Unsupervised Learning

- [ ] Clustering: K-Means, DBSCAN, Hierarchical clustering
- [ ] Dimensionality reduction: PCA, t-SNE
- [ ] Gaussian Mixture Models (GMM)
- [ ] Anomaly detection
- [ ] Applications of unsupervised learning

5. Model Selection & Evaluation

- [ ] Cross-validation (KFold, StratifiedKFold)

- [ ] Regression metrics: MSE, RMSE, MAE, R²
- [ ] Classification metrics: Accuracy, Precision, Recall, F1, ROC-AUC, Confusion Matrix
- [ ] Hyperparameter tuning: GridSearchCV, RandomizedSearchCV
- [ ] Learning & validation curves
- [ ] Saving/loading models (pickle, joblib)
- [ ] Bias-variance trade-off, overfitting vs underfitting

6. Pipelines & Production

- [ ] Full pipelines: preprocessing + model
- [ ] Feature engineering in pipelines
- [ ] Reproducibility (random_state)
- [ ] Handling large datasets (partial_fit)
- [ ] Parallelism (n_jobs)
- [ ] Model deployment basics

7. Advanced Topics & Best Practices

- [ ] Avoiding data leakage
- [ ] Feature importance & permutation importance
- [ ] Error analysis & debugging models
- [ ] Ensemble techniques: bagging, boosting, stacking
- [ ] Custom transformers & estimators
- [ ] Integrating with pandas, NumPy, Matplotlib
- [ ] Keeping up with sklearn updates

8. Practice Projects & Real-World Work

- [ ] End-to-end ML projects: raw data → model → results
- [ ] Classification & regression projects
- [ ] Unsupervised clustering or anomaly detection project
- [ ] Use pipelines + hyperparameter tuning
- [ ] Compare algorithms & analyze results
- [ ] Document findings and code cleanly