Applied Machine Learning Cheat Sheet

Data Preprocessing

- **Normalization:** $\mathbf{x} = \frac{\mathbf{x}-\min(\mathbf{x})}{\max(\mathbf{x})-\min(\mathbf{x})}$
- **Standardization:** $\mathbf{x} = \frac{\mathbf{x}-\mu}{\sigma}$
- **Handling Missing Values:**
  - Mean/Median Imputation
  - Mode Imputation
  - KNN Imputation

Feature Engineering

- **Polynomial Features:** $x_1, x_2, x_1^2, x_1x_2, x_2^2$
- **Interaction Features:** $x_1 \cdot x_2$
- **Log Transform:** $\log(x)$
- **Binning:** Discretize continuous variables

Model Selection

- **Linear Regression:** $y = \beta_0 + \beta_1 x + \epsilon$
- **Logistic Regression:** $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$
- **Decision Trees:** Recursive binary splitting
- **Random Forest:** Ensemble of decision trees
- **SVM:** Maximize margin between classes

Model Evaluation

- **Confusion Matrix:**
  - TP, FP, TN, FN
- **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$
- **Precision:** $\frac{TP}{TP+FP}$
- **Recall:** $\frac{TP}{TP+FN}$
- **F1 Score:** $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$
- **ROC Curve:** Plot of TPR vs. FPR

Hyperparameter Tuning

- **Grid Search:** Exhaustive search over parameter grid
- **Random Search:** Randomly sample parameters
- **Bayesian Optimization:** Model-based optimization
- **Cross-Validation:** k-Fold, Leave-One-Out