# Phase 2: Machine Learning Fundamentals

This document outlines the key topics, problems, and mini-projects to help you build a strong understanding of machine learning fundamentals. Completing these tasks will prepare you for more advanced ML concepts and practical applications.

## Topics and Problems to Solve

### 1. Supervised Learning

#### Key Concepts:

- Linear regression and logistic regression.

- Decision trees and random forests.

- Model evaluation metrics (accuracy, precision, recall, F1-score, ROC-AUC).

#### Problems:

1. Build a regression model to predict house prices based on features like square footage and number of bedrooms.

2. Create a classification model to detect spam emails based on a dataset of email text and labels.

3. Train a decision tree to classify customer churn (yes/no) based on user activity data.

4. Evaluate the performance of your models using precision, recall, and F1-score.

5. Create visualizations (e.g., confusion matrix) to better understand model performance.

---

### 2. Unsupervised Learning

#### Key Concepts:

- Clustering algorithms (K-Means).

- Dimensionality reduction techniques (Principal Component Analysis - PCA).

#### Problems:

1. Use K-Means to cluster customers based on purchase history data.

2. Apply PCA to reduce the dimensionality of a high-dimensional dataset and visualize the results in 2D.

3. Analyze the clusters formed by K-Means and describe their characteristics.

---

### 3. Data Preprocessing and Feature Engineering

#### Key Concepts:

- Handling missing data (imputation techniques like mean/median).

- Feature scaling (normalization and standardization).

- Encoding categorical variables (One-Hot Encoding, Label Encoding).

#### Problems:

1. Handle missing values in a dataset using mean and median imputation.

2. Apply MinMaxScaler and StandardScaler to normalize and standardize features.

3. Encode categorical variables (e.g., city names, product categories) in a dataset.

4. Create new features from existing ones (e.g., total sales = price \* quantity).

---

### 4. Model Evaluation and Hyperparameter Tuning

#### Key Concepts:

- Train/Test split and cross-validation.

- Overfitting and underfitting.

- Hyperparameter tuning techniques (GridSearchCV, RandomizedSearchCV).

#### Problems:

1. Split a dataset into training and test sets, then evaluate model performance on both.

2. Use k-fold cross-validation to validate a model's performance.

3. Tune hyperparameters of a random forest model to improve accuracy.

4. Analyze how model performance changes with different hyperparameter values.

---

## Mini-Projects

### Project 1: Predicting House Prices

- \*\*Objective\*\*: Build and evaluate a regression model.

- \*\*Dataset\*\*: A housing dataset with columns like `square\_footage`, `bedrooms`, `location`, and `price`.

- \*\*Tasks\*\*:

1. Preprocess the dataset (handle missing values, encode categorical variables).

2. Train a linear regression model.

3. Evaluate the model using RMSE and R^2 metrics.

### Project 2: Customer Churn Prediction

- \*\*Objective\*\*: Create and deploy a classification model to predict customer churn.

- \*\*Dataset\*\*: A dataset with customer activity features and a `churn` label.

- \*\*Tasks\*\*:

1. Preprocess the dataset (normalize numerical features, encode categorical variables).

2. Train a logistic regression and decision tree model.

3. Evaluate model performance and select the best-performing model.

### Project 3: Clustering Customer Segments

- \*\*Objective\*\*: Use unsupervised learning to segment customers into meaningful groups.

- \*\*Dataset\*\*: A dataset with customer attributes like `age`, `income`, `purchase\_frequency`.

- \*\*Tasks\*\*:

1. Apply K-Means clustering and determine the optimal number of clusters using the elbow method.

2. Visualize clusters using scatter plots or PCA.

3. Describe the characteristics of each cluster.

---

## Tools

- \*\*Python\*\*: Core programming language.

- \*\*Scikit-learn\*\*: For machine learning models and preprocessing.

- \*\*Pandas\*\*: For data manipulation.

- \*\*Matplotlib/Seaborn\*\*: For visualizations.

---

## Outcomes

By completing these tasks, you will:

1. Understand the basics of supervised and unsupervised learning.

2. Build and evaluate models for regression and classification problems.

3. Learn to preprocess data and engineer features effectively.

4. Complete at least \*\*3 mini-projects\*\* to showcase your ML skills.

---

Let me know if you need further guidance or additional resources to complete these tasks!