

# Chapter 38

## Scikit-Learn

Comprehensive Notes

### Contents

<b>1</b>	<b>Introduction to Scikit-Learn</b>	<b>3</b>
<b>2</b>	<b>Data Representation in Scikit-Learn</b>	<b>3</b>
2.1	The Features Matrix and Target Array . . . . .	3
2.2	Example with the Iris Dataset . . . . .	4
2.3	Preparing Data for Scikit-Learn . . . . .	5
<b>3</b>	<b>The Estimator API</b>	<b>6</b>
3.1	The Standard Workflow Pattern . . . . .	7
<b>4</b>	<b>Supervised Learning Examples</b>	<b>7</b>
4.1	Example 1: Simple Linear Regression . . . . .	7
4.1.1	Step 1: Create some sample data . . . . .	7
4.1.2	Step 2: Choose a model class . . . . .	8
4.1.3	Step 3: Instantiate the model with hyperparameters . . . . .	9
4.1.4	Step 4: Arrange data into features matrix and target vector . . . . .	9
4.1.5	Step 5: Fit the model to the data . . . . .	9
4.1.6	Step 6: Use the model to make predictions . . . . .	10
4.2	Example 2: Iris Classification . . . . .	11
4.2.1	Step 1: Split data into training and testing sets . . . . .	11
4.2.2	Step 2-3: Choose and instantiate a model . . . . .	12
4.2.3	Step 4: Fit the model to the training data . . . . .	12
4.2.4	Step 5: Make predictions on the test data . . . . .	12
4.2.5	Step 6: Evaluate model performance . . . . .	13
<b>5</b>	<b>Unsupervised Learning Examples</b>	<b>14</b>
5.1	Example 1: Dimensionality Reduction with PCA . . . . .	14
5.2	Example 2: Clustering with Gaussian Mixture Models . . . . .	16
<b>6</b>	<b>Application: Exploring Handwritten Digits</b>	<b>18</b>
6.1	Loading and Visualizing the Digits Data . . . . .	18
6.2	Preparing the Data for Machine Learning . . . . .	19
6.3	Dimensionality Reduction for Visualization . . . . .	20
6.4	Building a Classification Model . . . . .	21
6.5	Creating and Visualizing the Confusion Matrix . . . . .	21
6.6	Visualizing Correct and Incorrect Predictions . . . . .	22

<b>7</b>	<b>Step-by-Step Process: A Comprehensive Guide</b>	<b>24</b>
7.1	1. Understand the Problem . . . . .	24
7.2	2. Explore and Prepare the Data . . . . .	24
7.3	3. Select a Model . . . . .	25
7.4	4. Train and Tune the Model . . . . .	26
7.5	5. Evaluate the Model . . . . .	27
7.6	6. Deploy and Monitor the Model . . . . .	28
<b>8</b>	<b>Summary</b>	<b>28</b>

# 1 Introduction to Scikit-Learn

Scikit-Learn is one of the most widely used Python libraries for machine learning, providing efficient implementations of a large number of common algorithms. Its popularity stems from several key features:

## What Makes Scikit-Learn Special?

- **Clean, uniform API:** Once you learn the basic pattern, all models follow the same workflow
- **Comprehensive documentation:** Well-explained algorithms with theory and examples
- **Consistent interface:** Makes switching between models straightforward
- **Optimized implementations:** Efficient versions of common algorithms

This chapter provides an overview of the Scikit-Learn API, focusing on:

1. How data is represented within Scikit-Learn
2. The Estimator API that forms the foundation of all models
3. Practical examples demonstrating supervised and unsupervised learning
4. A deeper application using handwritten digits classification

A solid understanding of these API elements will form the foundation for exploring machine learning algorithms and approaches in your own projects.

## 2 Data Representation in Scikit-Learn

Machine learning is fundamentally about creating models from data. Therefore, before diving into algorithms, we need to understand how data should be structured for use with Scikit-Learn.

### 2.1 The Features Matrix and Target Array

#### Scikit-Learn's Data Structure

- **Features Matrix (X):** Two-dimensional array with shape `[n_samples, n_features]`
- **Target Array (y):** One-dimensional array with length `n_samples`
- **Common data containers:** NumPy arrays, Pandas DataFrames, SciPy sparse matrices

The best way to think about data within Scikit-Learn is in terms of tables:

- **Rows** represent individual samples (observations)
- **Columns** represent features (variables or attributes)

## 2.2 Example with the Iris Dataset

Let's examine the classic Iris dataset as an example:

```
1 import seaborn as sns
2
3 # Load the dataset
4 iris = sns.load_dataset('iris')
5
6 # Display first few rows
7 iris.head()
```

Listing 1: Loading the Iris dataset with Seaborn

Output:

	sepal.length	sepal.width	petal.length	petal.width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In this dataset:

- Each row represents one flower (a sample)
- The first four columns are features (measurements)
- The last column (species) is what we typically want to predict (the target)

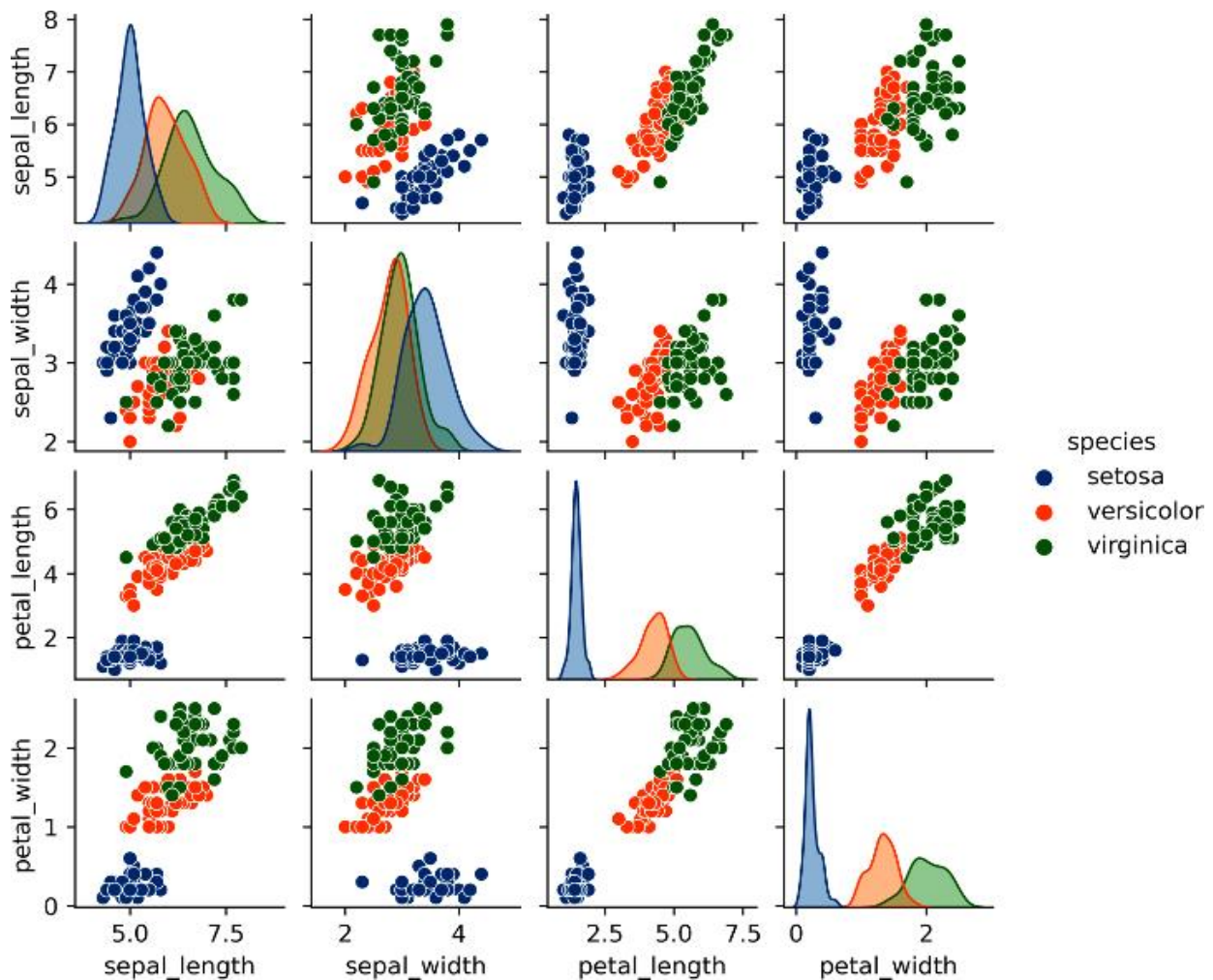
### Understanding the Target Array

The target array is the quantity we want to predict from the features. In statistical terms, it's the dependent variable. For the Iris dataset, if we want to predict the species based on measurements, then the species column would be our target, and the measurements would be our features.

We can visualize the data with Seaborn's pairplot to see relationships between features:

```
1 %matplotlib inline
2 import seaborn as sns
3 sns.pairplot(iris, hue='species', height=1.5)
```

Listing 2: Visualizing the Iris dataset



Pairplot visualization of the Iris dataset showing relationships between features, colored by species.

## 2.3 Preparing Data for Scikit-Learn

For use in Scikit-Learn, we need to extract the features matrix and target array:

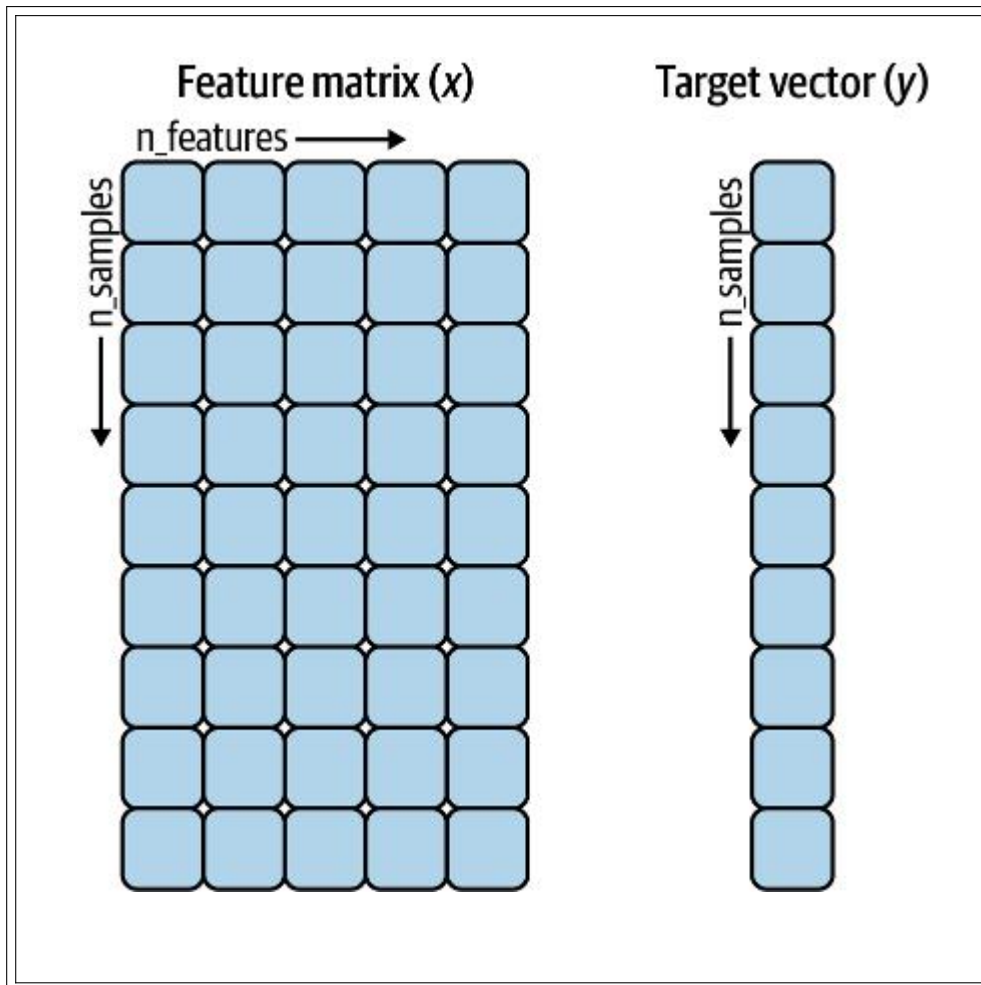
```

1 # Extract features (all columns except 'species')
2 X_iris = iris.drop('species', axis=1)
3 print(X_iris.shape) # (150, 4)
4
5 # Extract target (just the 'species' column)
6 y_iris = iris['species']
7 print(y_iris.shape) # (150,)

```

Listing 3: Extracting features and target from Iris dataset

The expected layout of features and target looks like this:



**Figure 2:**

Scikit-Learn's data layout showing the features matrix (X) and target vector (y).

### Pro Tip

When working with real-world data, you'll typically need to:

1. Clean the data (handle missing values, outliers, etc.)
2. Convert categorical variables into numerical form
3. Scale or normalize numerical features
4. Split into training and testing sets

Scikit-Learn provides utilities for all these preprocessing steps.

## 3 The Estimator API

The Scikit-Learn API is designed with several guiding principles that make it particularly user-friendly:

### Scikit-Learn API Design Principles

- **Consistency:** All objects share a common interface with consistent documentation
- **Inspection:** All parameter values are exposed as public attributes

- **Limited object hierarchy:** Algorithms are represented by Python classes; datasets use standard formats
- **Composition:** Machine learning tasks can be expressed as sequences of more fundamental algorithms
- **Sensible defaults:** When parameters are required, appropriate default values are defined

### 3.1 The Standard Workflow Pattern

Regardless of the specific algorithm, Scikit-Learn follows a consistent pattern:



Let's break down each step:

1. **Choose a class of model:** Import the appropriate estimator class from Scikit-Learn
2. **Choose model hyperparameters:** Instantiate the class with desired values
3. **Arrange data:** Format your data as features matrix  $X$  and target vector  $y$
4. **Fit the model:** Call the `fit()` method with your training data
5. **Apply the model:** Use the model for predictions or transformations

#### Insight

##### Supervised vs. Unsupervised Learning:

- **Supervised learning:** We predict labels using the `predict()` method (classification, regression)
- **Unsupervised learning:** We transform or infer properties using methods like `transform()` or `predict()` (clustering, dimensionality reduction)

## 4 Supervised Learning Examples

### 4.1 Example 1: Simple Linear Regression

Let's step through a complete example of linear regression following our workflow.

#### 4.1.1 Step 1: Create some sample data

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 # Set random seed for reproducibility
5 rng = np.random.RandomState(42)
6
7 # Create data with a linear relationship plus some noise
8 x = 10 * rng.rand(50)          # 50 random x values between 0
   and 10
```

```

9 y = 2 * x - 1 + rng.randn(50)          # y = 2x - 1 + noise
10
11 # Visualize the data
12 plt.figure(figsize=(8, 6))
13 plt.scatter(x, y)
14 plt.xlabel('x')
15 plt.ylabel('y')
16 plt.title('Data for Linear Regression')
17 plt.show()

```

Listing 4: Creating sample data for linear regression

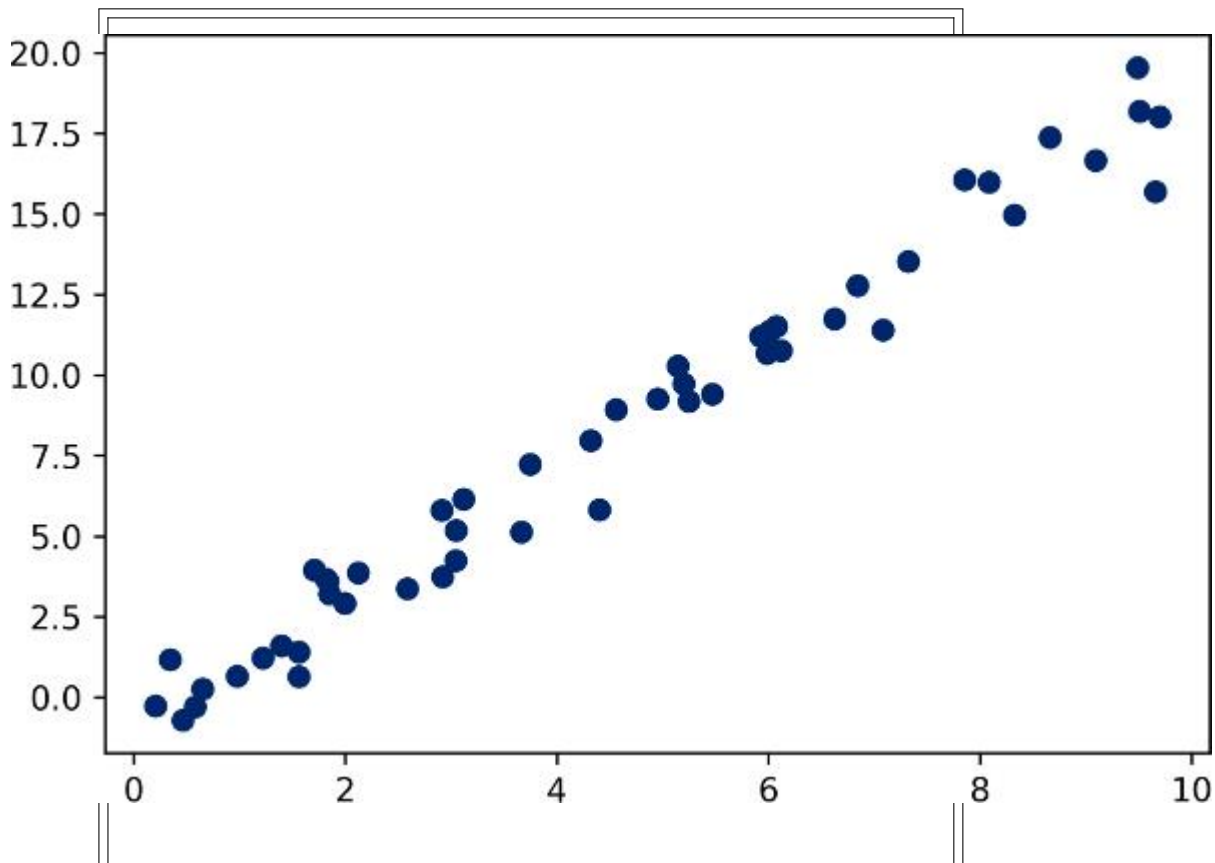


Figure 3: Scatter plot of our sample data showing an approximately linear relationship.

#### 4.1.2 Step 2: Choose a model class

```

1 from sklearn.linear_model import LinearRegression

```

Listing 5: Importing the LinearRegression model

#### Insight

Scikit-Learn organizes models into submodules by task:

- `sklearn.linear_model`: Linear models (Linear/Logistic Regression, etc.)
- `sklearn.ensemble`: Ensemble methods (Random Forests, Gradient Boosting, etc.)
- `sklearn.tree`: Decision Trees



- `sklearn.cluster`: Clustering algorithms
- `sklearn.decomposition`: Dimensionality reduction

#### 4.1.3 Step 3: Instantiate the model with hyperparameters

```

1 # Create a linear regression model with the intercept
2 model = LinearRegression(fit_intercept=True)
3
4 # Examine the model
5 print(model) # Output: LinearRegression()

```

Listing 6: Instantiating the LinearRegression model

### Understanding Hyperparameters

**Hyperparameters** are parameters set before learning begins, distinguishing them from model parameters that are learned during training.

For `LinearRegression`, hyperparameters include:

- `fit_intercept`: Whether to calculate the intercept (default=True)
- `normalize`: Whether to normalize features (deprecated)
- `copy_X`: Whether to copy X (default=True)
- `n_jobs`: Number of jobs for computation (default=None)
- `positive`: Whether to force coefficients to be positive (default=False)

Other models may have many more hyperparameters that significantly impact performance.

#### 4.1.4 Step 4: Arrange data into features matrix and target vector

```

1 # Reshape x to be a 2D array (required by Scikit-Learn)
2 X = x[:, np.newaxis] # From shape (50,) to shape (50, 1)
3
4 print(f"X shape: {X.shape}") # Output: X shape: (50, 1)
5 print(f"y shape: {y.shape}") # Output: y shape: (50,)

```

Listing 7: Preparing the data for Scikit-Learn

### Warning

Scikit-Learn requires the features to be in a 2D array, even if there's only one feature. If you have a 1D array, you must reshape it using `np.newaxis` or `.reshape(-1, 1)`. The target variable can stay as a 1D array.

#### 4.1.5 Step 5: Fit the model to the data

```

1 # Train the model on our data

```

```

2 model.fit(X, y)
3
4 # After fitting, the model has learned parameters
5 print(f"Coefficient (slope): {model.coef_[0]:.4f}")           # Should be
6     close to 2
7 print(f"Intercept: {model.intercept_[0]:.4f}")               # Should be
8     close to -1

```

Listing 8: Fitting the linear regression model

### Insight

After calling `fit()`, Scikit-Learn models store the learned parameters with a trailing underscore (e.g., `coef_`, `intercept_`). These represent the parameters that were learned from the data.

For linear regression, these parameters correspond to the slope and intercept of the line. Comparing with our data generation ( $y = 2x - 1 + \text{noise}$ ), we should see values close to 2 and -1.

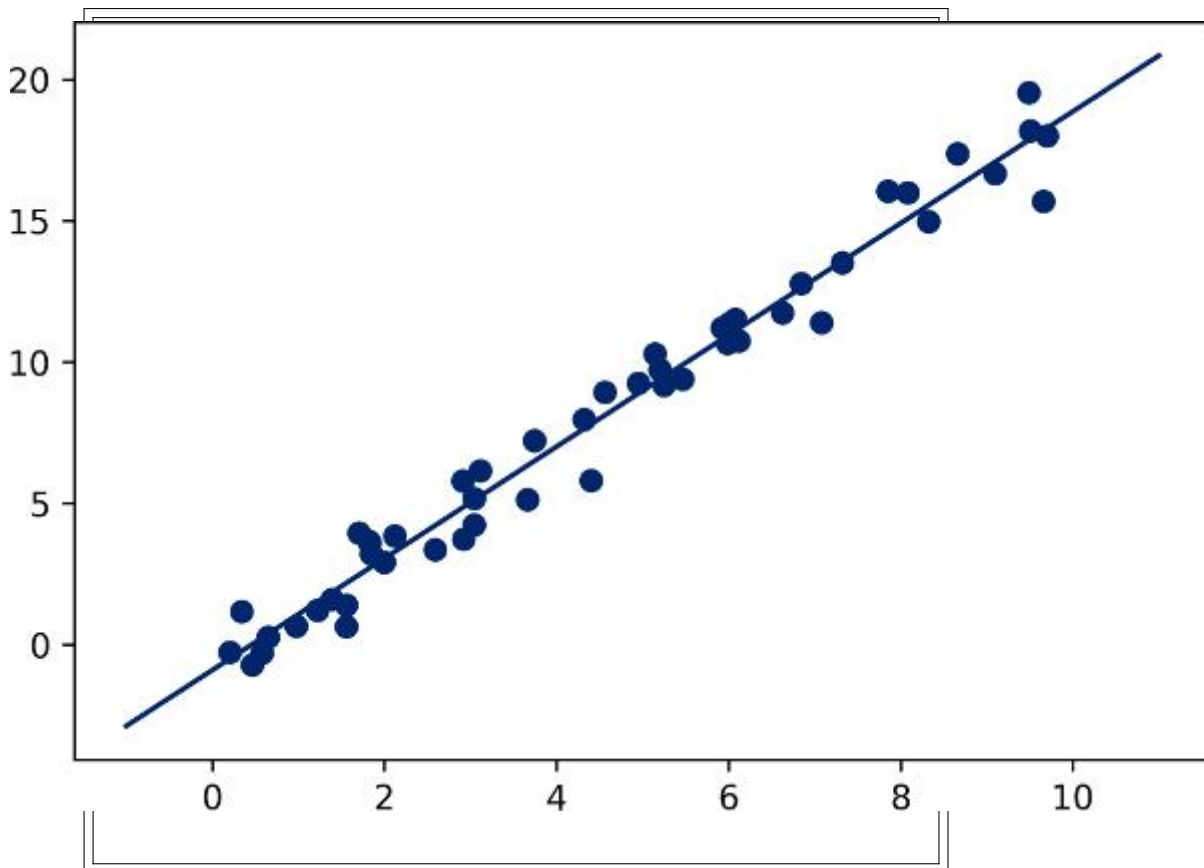
#### 4.1.6 Step 6: Use the model to make predictions

```

1 # Create a range of x values for prediction
2 xfit = np.linspace(-1, 11, 100) # 100 points from -1 to 11
3 Xfit = xfit[:, np.newaxis]       # Reshape for Scikit-Learn
4
5 # Predict y values using our trained model
6 yfit = model.predict(Xfit)
7
8 # Visualize the results
9 plt.figure(figsize=(8, 6))
10 plt.scatter(x, y, alpha=0.7)      # Original data points
11 plt.plot(xfit, yfit, color='red', linewidth=2) # Regression line
12 plt.xlabel('x')
13 plt.ylabel('y')
14 plt.title('Linear Regression Fit')
15 plt.show()

```

Listing 9: Making predictions with the trained model



**Figure 4:** Linear regression fit (red line) to our data points.

### Pro Tip

When evaluating regression models, common metrics include:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- $R^2$  Score (coefficient of determination)

Scikit-Learn provides these metrics in the `sklearn.metrics` module:

```
1 from sklearn.metrics import mean_squared_error, r2_score
2 mse = mean_squared_error(y_true, y_pred)
3 r2 = r2_score(y_true, y_pred) # 1.0 is perfect prediction
```

## 4.2 Example 2: Iris Classification

Now let's apply our pattern to a classification problem using the Iris dataset we saw earlier.

### 4.2.1 Step 1: Split data into training and testing sets

For proper evaluation, we split our data into separate sets for training and testing:

```
1 from sklearn.model_selection import train_test_split
```

```

2
3 # Split data into train and test sets (75% train, 25% test)
4 X_train, X_test, y_train, y_test = train_test_split(
5     X_iris, y_iris, random_state=1)
6
7 print(f"Training set size: {X_train.shape[0]} samples")
8 print(f"Testing set size: {X_test.shape[0]} samples")

```

Listing 10: Splitting data into training and testing sets

### The Importance of Train-Test Splits

Splitting data into separate training and testing sets is crucial for evaluating model performance on unseen data. This helps detect overfitting (when a model performs well on training data but poorly on new data).

The `random_state` parameter ensures reproducible splits. Setting a specific value means you'll get the same split every time you run the code.

#### 4.2.2 Step 2-3: Choose and instantiate a model

For this task, we'll use Gaussian Naive Bayes, a simple but effective classifier:

```

1 from sklearn.naive_bayes import GaussianNB
2
3 # Create a Gaussian Naive Bayes model
4 model = GaussianNB()
5
6 # Learn more about the model parameters
7 print(model) # GaussianNB()

```

Listing 11: Creating a Gaussian Naive Bayes model

### Pro Tip

Gaussian Naive Bayes assumes:

- Features are independent of each other (hence "naive")
- Feature values follow a Gaussian (normal) distribution for each class

Despite these simplifying assumptions, it often performs well and is very fast to train and predict, making it a good baseline classifier.

#### 4.2.3 Step 4: Fit the model to the training data

```

1 # Train the model on the training data
2 model.fit(X_train, y_train)

```

Listing 12: Training the Naive Bayes model

#### 4.2.4 Step 5: Make predictions on the test data

```

1 # Predict species for test data
2 y_pred = model.predict(X_test)
3
4 # Show first few predictions vs true values
5 for i in range(5):
6     print(f"Sample {i}: True={y_test.iloc[i]}, Predicted={y_pred[i]}")

```

Listing 13: Making predictions with the trained model

#### 4.2.5 Step 6: Evaluate model performance

```

1 from sklearn.metrics import accuracy_score, classification_report
2
3 # Calculate accuracy
4 accuracy = accuracy_score(y_test, y_pred)
5 print(f"Accuracy: {accuracy:.4f}")
6
7 # Detailed classification report
8 print("\nClassification Report:")
9 print(classification_report(y_test, y_pred))

```

Listing 14: Evaluating classification accuracy

#### Understanding Classification Metrics

Beyond accuracy, it's important to consider:

- **Precision:** The proportion of positive identifications that were actually correct
- **Recall:** The proportion of actual positives that were identified correctly
- **F1-score:** The harmonic mean of precision and recall
- **Confusion matrix:** Table showing correct and incorrect classifications per class

The `classification_report` function provides these metrics for each class.

```

1 from sklearn.metrics import confusion_matrix
2 import seaborn as sns
3
4 # Compute the confusion matrix
5 cm = confusion_matrix(y_test, y_pred)
6
7 # Visualize the confusion matrix
8 plt.figure(figsize=(8, 6))
9 sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
10             xticklabels=model.classes_,
11             yticklabels=model.classes_)
12 plt.xlabel('Predicted Label')
13 plt.ylabel('True Label')
14 plt.title('Confusion Matrix')
15 plt.show()

```

Listing 15: Creating a confusion matrix



**Figure 5:** Confusion matrix showing the performance of our classifier.

## 5 Unsupervised Learning Examples

### 5.1 Example 1: Dimensionality Reduction with PCA

Principal Component Analysis (PCA) is a technique for reducing the dimensionality of data while preserving as much variance as possible.

```
1 from sklearn.decomposition import PCA
2
3 # Create a PCA model with 2 components
4 pca = PCA(n_components=2)
5
6 # Fit the model to the data
7 pca.fit(X_iris)
8
9 # Transform the data to the new 2D space
10 X_2D = pca.transform(X_iris)
11
12 print(f"Original shape: {X_iris.shape}") # (150, 4)
13 print(f"Reduced shape: {X_2D.shape}")    # (150, 2)
14
15 # How much variance is explained by each component?
16 explained_variance = pca.explained_variance_ratio_
17 print(f"Explained variance ratio: {explained_variance}")
18 print(f"Total variance explained: {sum(explained_variance):.2f}")
```

Listing 16: Applying PCA to the Iris dataset

## What is PCA?

PCA finds directions (principal components) in the feature space along which the data varies the most. By projecting the data onto these components, we can reduce dimensionality while preserving as much information as possible.

The `explained_variance_ratio_` tells us how much of the total variance in the data is explained by each principal component.

```
1 # Add PCA results to our DataFrame for visualization
2 iris_pca = iris.copy()
3 iris_pca['PCA1'] = X_2D[:, 0]
4 iris_pca['PCA2'] = X_2D[:, 1]
5
6 # Plot the 2D projection with colors by species
7 plt.figure(figsize=(10, 8))
8 sns.scatterplot(x='PCA1', y='PCA2', hue='species',
9                 data=iris_pca, palette='viridis', s=100, alpha=0.7)
10 plt.title('PCA of Iris Dataset')
11 plt.xlabel('First Principal Component')
12 plt.ylabel('Second Principal Component')
13 plt.legend(title='Species')
14 plt.show()
```

Listing 17: Visualizing PCA results

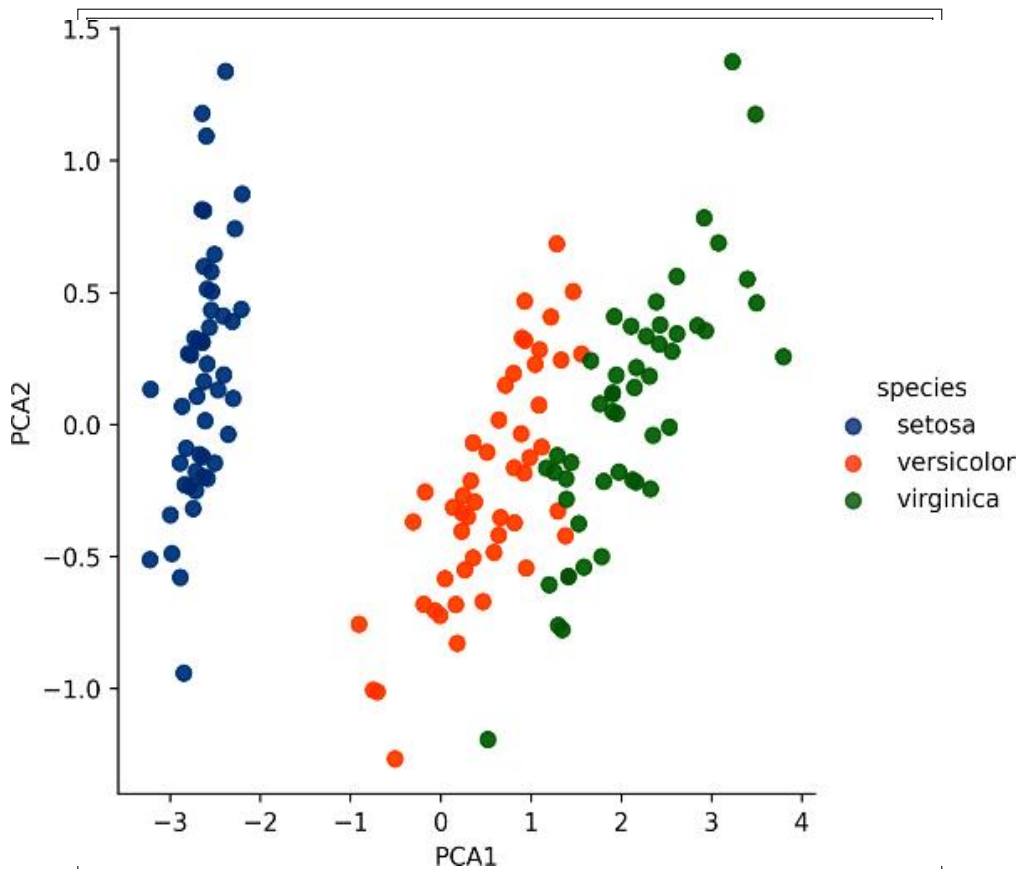


Figure 6: Iris data projected onto its first two principal components, colored by species.

### Pro Tip

When we see good separation of classes in the reduced dimensions (as in this case), it suggests that a classification algorithm should work well. The fact that the species form distinct clusters without us telling PCA about the class labels is a good sign!

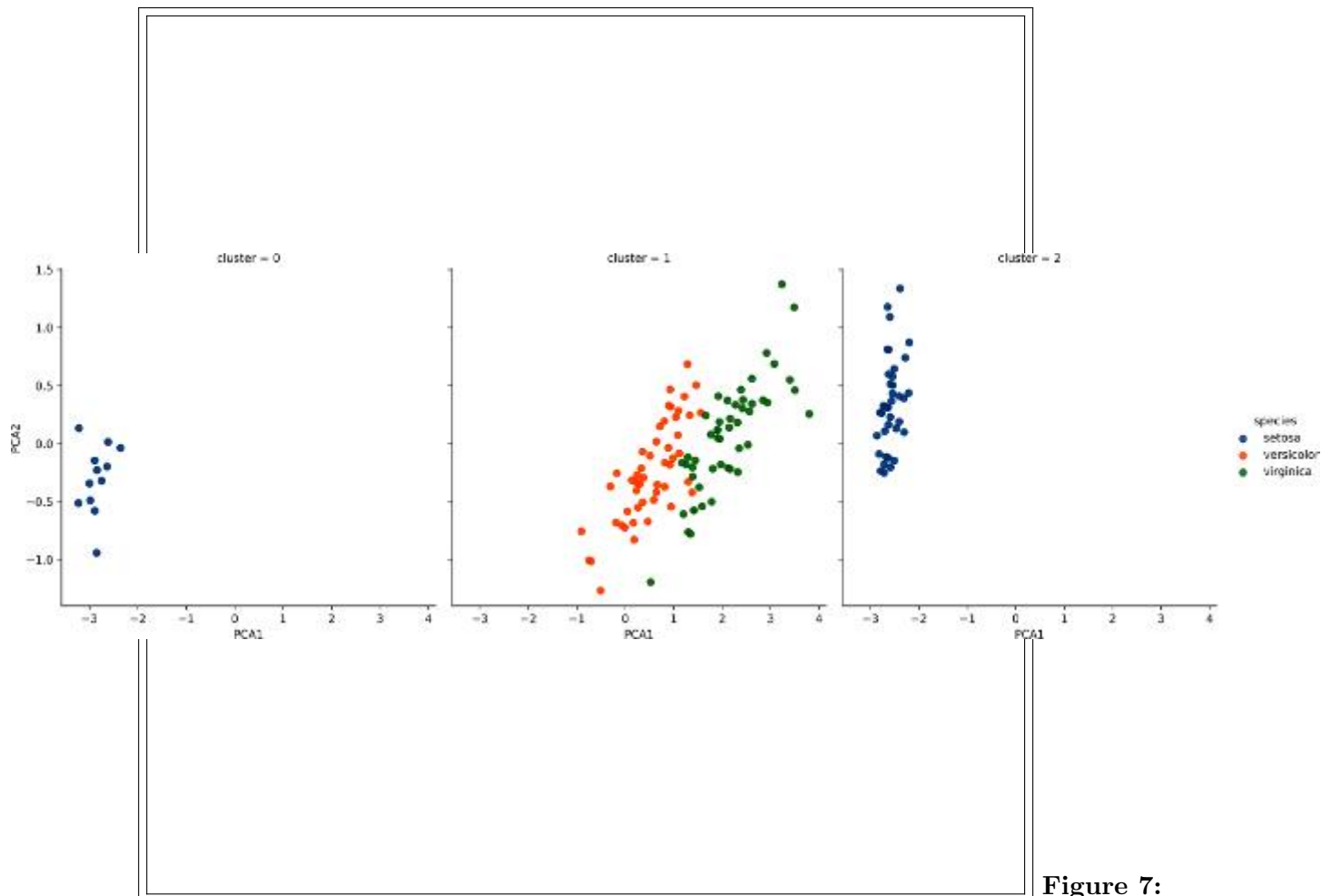
## 5.2 Example 2: Clustering with Gaussian Mixture Models

Clustering is an unsupervised learning technique that groups similar data points together.

```
1 from sklearn.mixture import GaussianMixture
2
3 # Create a Gaussian Mixture Model with 3 components
4 gmm = GaussianMixture(n_components=3, covariance_type='full',
5     random_state=42)
6
7 # Fit the model to the data
8 gmm.fit(X_iris)
9
10 # Predict the cluster for each data point
11 clusters = gmm.predict(X_iris)
12
13 # Add cluster information to our DataFrame
14 iris_clustered = iris.copy()
15 iris_clustered['cluster'] = clusters
16
17 # Visualize the clusters using the PCA projection from before
18 plt.figure(figsize=(12, 5))
19
20 # Plot 1: Color by predicted cluster
21 plt.subplot(1, 2, 1)
22 sns.scatterplot(x='PCA1', y='PCA2', hue='cluster',
23     data=iris_clustered, palette='Set1', s=80, alpha=0.7)
24 plt.title('GMM Clustering Results')
25
26 # Plot 2: Color by actual species
27 plt.subplot(1, 2, 2)
28 sns.scatterplot(x='PCA1', y='PCA2', hue='species',
29     data=iris_clustered, palette='viridis', s=80, alpha=0.7)
30 plt.title('Actual Species')
31
32 plt.tight_layout()
33 plt.show()
```

Listing 18: Clustering the Iris data with Gaussian Mixture Models





**Figure 7:**  
Comparison of GMM clustering results (left) with actual species (right).

### Gaussian Mixture Models (GMM)

A Gaussian Mixture Model represents data as a mixture of several Gaussian distributions. GMM:

- Allows for soft clustering (points can belong partially to multiple clusters)
- Can capture clusters of different sizes and orientations
- Provides a probability distribution over the clusters
- Is more flexible than simpler algorithms like K-means

Let's evaluate how well our clusters correspond to the actual species:

```
1 # Cross-tabulation of clusters vs actual species
2 crosstab = pd.crosstab(
3     iris_clustered['species'],
4     iris_clustered['cluster'],
5     rownames=['Species'],
6     colnames=['Cluster'])
7
8 print(crosstab)
```

Listing 19: Evaluating clustering with a cross-tabulation

### Pro Tip

Unlike supervised learning, there's no single "correct" way to evaluate clustering results:

- **When ground truth labels exist:** Use metrics like adjusted Rand index, normalized mutual information
- **Without ground truth:** Use metrics like silhouette score, Davies-Bouldin index, or inertia

Scikit-Learn provides these metrics in the `sklearn.metrics.cluster` module.

## 6 Application: Exploring Handwritten Digits

Let's apply what we've learned to a more complex problem: recognizing handwritten digits. This is a subset of optical character recognition (OCR) and is a classic problem in machine learning.

### 6.1 Loading and Visualizing the Digits Data

```
1 from sklearn.datasets import load_digits
2
3 # Load the digits dataset
4 digits = load_digits()
5
6 # Examine the dataset structure
7 print(f"Data shape: {digits.images.shape}")           # (1797, 8, 8)
8 print(f"Target shape: {digits.target.shape}")         # (1797,)
9 print(f"Classes: {np.unique(digits.target)}")         # [0 1 2 3 4 5 6
10               7 8 9]
```

Listing 20: Loading the digits dataset

The digits dataset contains:

- 1,797 samples (handwritten digits)
- Each digit is represented as an 8×8 pixel image
- 10 classes (digits 0-9)

Let's visualize some of these digits:

```
1 import matplotlib.pyplot as plt
2
3 # Create a figure with subplots
4 fig, axes = plt.subplots(4, 10, figsize=(12, 5),
5                           subplot_kw={'xticks': [], 'yticks': []},
6                           gridspec_kw=dict(hspace=0.1, wspace=0.1))
7
8 # Display the first 40 digits
9 for i, ax in enumerate(axes.flat):
10     if i < 40: # Only fill the first 40 slots
11         ax.imshow(digits.images[i], cmap='binary')
12         ax.text(0.05, 0.05, str(digits.target[i]),
13                transform=ax.transAxes, color='green')
```

```

14
15 plt.show()

```

Listing 21: Visualizing sample digits

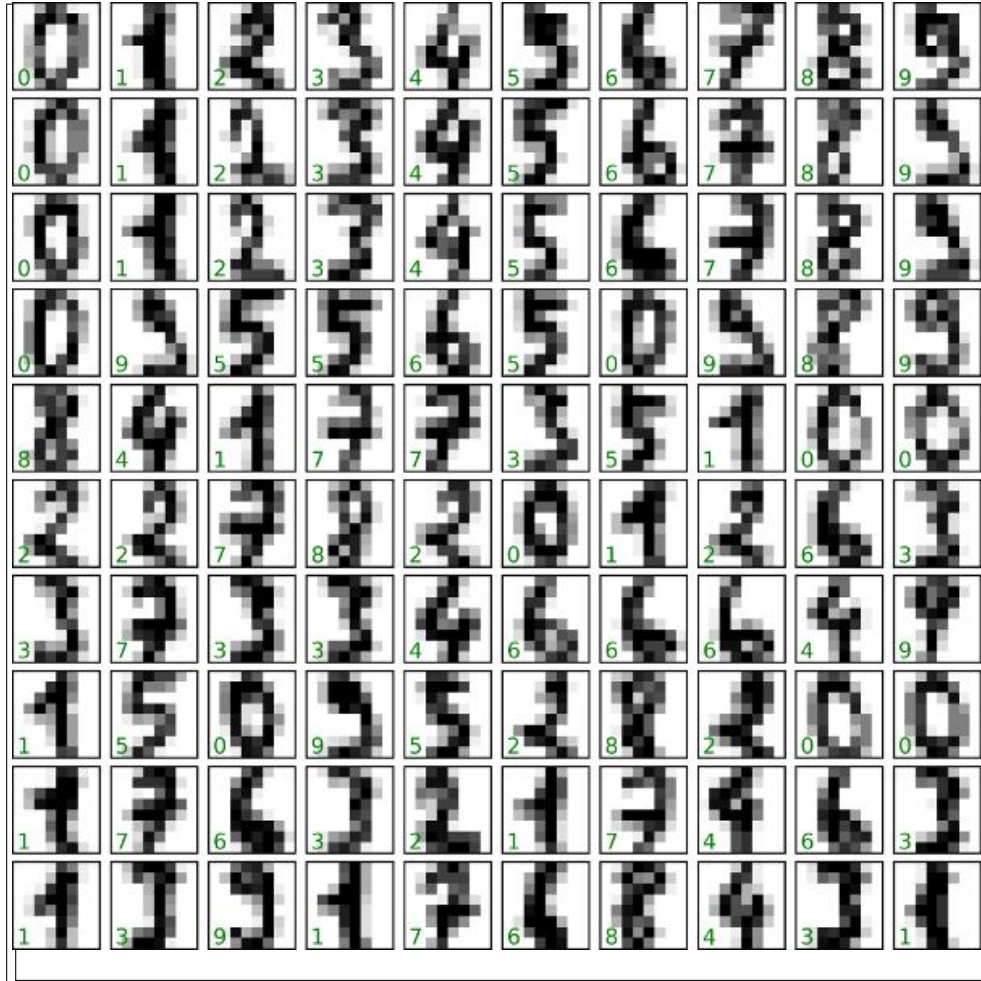


Figure 8: Sample digits from the dataset with their labels.

## 6.2 Preparing the Data for Machine Learning

For Scikit-Learn, we need to reshape the  $8 \times 8$  images into feature vectors:

```

1 # Flatten the images into feature vectors
2 X = digits.data # digits.data is already flattened: (1797, 64)
3 y = digits.target
4
5 print(f"Features shape: {X.shape}") # (1797, 64)
6 print(f"Target shape: {y.shape}") # (1797,)

```

Listing 22: Preparing the digits data for Scikit-Learn

### Insight

Each  $8 \times 8$  image is flattened into a 64-element vector, where each element represents the intensity of one pixel (from 0 to 16). This transforms our 3D array of images into a 2D features matrix suitable for Scikit-Learn.

### 6.3 Dimensionality Reduction for Visualization

Let's use a dimensionality reduction technique to visualize the high-dimensional data:

```
1 from sklearn.manifold import Isomap
2
3 # Create and fit the Isomap model
4 iso = Isomap(n_components=2)
5 iso.fit(X)
6
7 # Transform the data to 2D
8 X_projected = iso.transform(X)
9
10 # Plot the projected data
11 plt.figure(figsize=(10, 8))
12 plt.scatter(X_projected[:, 0], X_projected[:, 1],
13             c=y, cmap='viridis',
14             edgecolor='none', alpha=0.7, s=60)
15
16 plt.colorbar(label='Digit Label', ticks=range(10))
17 plt.clim(-0.5, 9.5)
18 plt.title('Isomap Projection of Handwritten Digits')
19 plt.xlabel('First Component')
20 plt.ylabel('Second Component')
21 plt.show()
```

Listing 23: Applying Isomap for dimensionality reduction

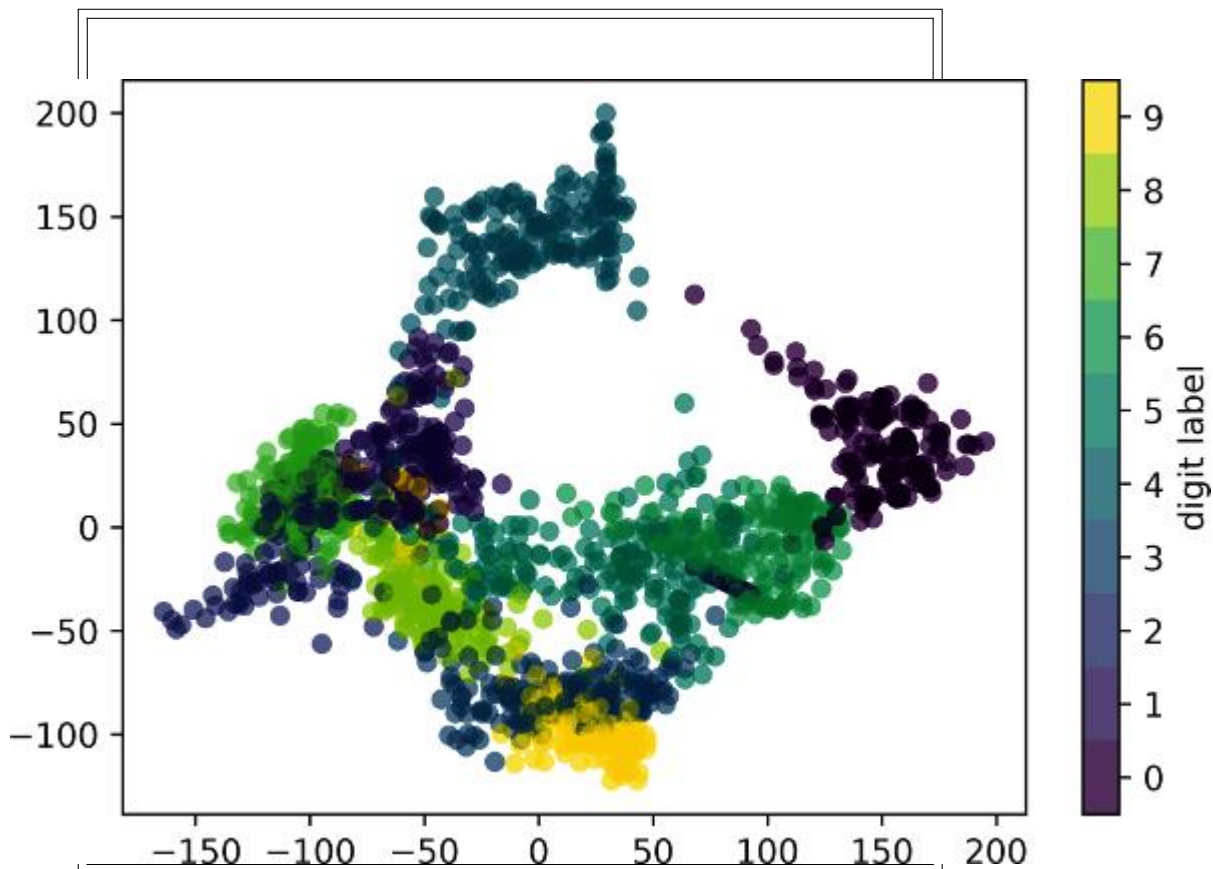


Figure 9: Handwritten digits projected to 2D using Isomap.

## Isomap vs. PCA

**Isomap** is a manifold learning technique that seeks to preserve the geodesic distances between points, which can capture non-linear structures in the data. In contrast, **PCA** preserves linear relationships.

Isomap is particularly good for data that lies on a non-linear manifold, such as images of handwritten digits with varied styles.

## 6.4 Building a Classification Model

Now, let's build a classifier to recognize the digits:

```
1 from sklearn.model_selection import train_test_split
2 from sklearn.naive_bayes import GaussianNB
3 from sklearn.metrics import accuracy_score, confusion_matrix
4
5 # Split data into training and testing sets
6 X_train, X_test, y_train, y_test = train_test_split(
7     X, y, test_size=0.25, random_state=42)
8
9 # Create and train a Gaussian Naive Bayes model
10 model = GaussianNB()
11 model.fit(X_train, y_train)
12
13 # Make predictions on the test set
14 y_pred = model.predict(X_test)
15
16 # Evaluate accuracy
17 accuracy = accuracy_score(y_test, y_pred)
18 print(f"Model accuracy: {accuracy:.4f}")
```

Listing 24: Training a classifier on the digits dataset

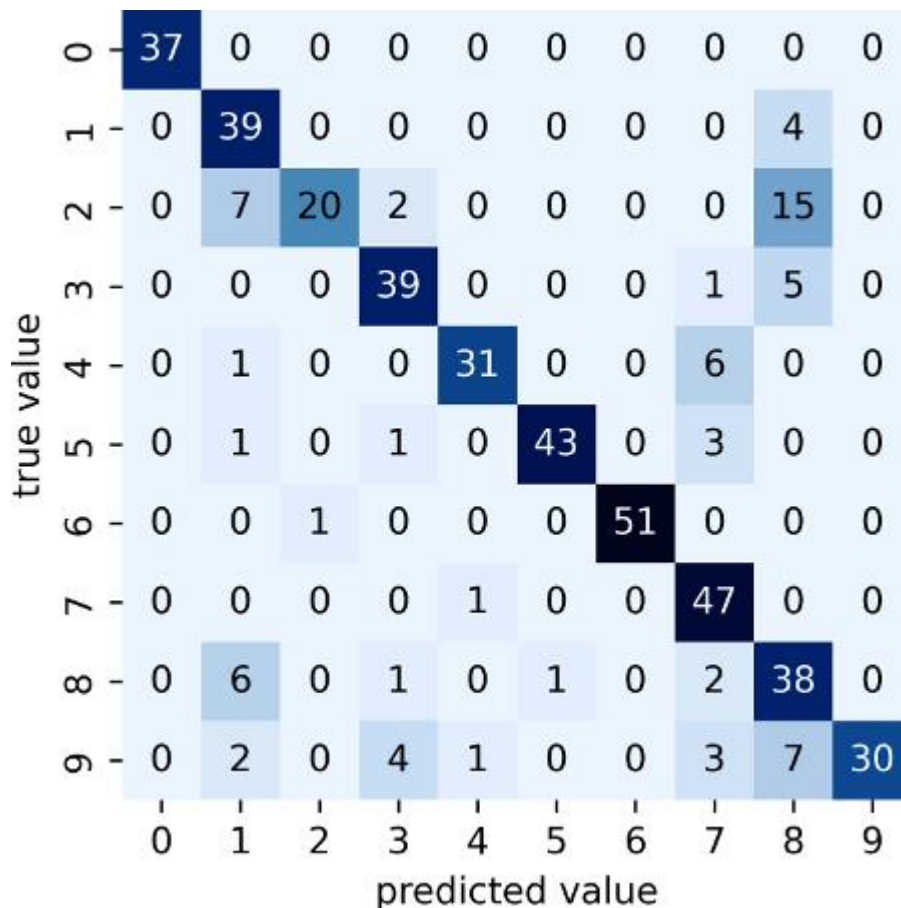
## Insight

Even with a simple model like Gaussian Naive Bayes, we can get quite good accuracy on digit recognition. This is because the digit classes form relatively distinct clusters in the feature space, as we saw in the Isomap visualization.

## 6.5 Creating and Visualizing the Confusion Matrix

```
1 # Compute the confusion matrix
2 cm = confusion_matrix(y_test, y_pred)
3
4 # Create a nicer visualization
5 plt.figure(figsize=(10, 8))
6 sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
7             xticklabels=range(10),
8             yticklabels=range(10))
9 plt.xlabel('Predicted Digit')
10 plt.ylabel('True Digit')
11 plt.title('Confusion Matrix for Digit Recognition')
12 plt.show()
```

Listing 25: Visualizing the confusion matrix



**Figure 10:** Confusion matrix for digit classification.

### Pro Tip

The confusion matrix is a powerful tool for understanding classification errors:

- Rows represent the true digit
- Columns represent the predicted digit
- Diagonal elements are correct classifications
- Off-diagonal elements are mistakes

Examining the confusion matrix can reveal patterns in misclassifications, like which digits are commonly confused with each other.

## 6.6 Visualizing Correct and Incorrect Predictions

Let's visualize some of our model's predictions, highlighting correct and incorrect classifications:

```

1 # Reshape test data back to images
2 test_images = X_test.reshape(-1, 8, 8)
3
4 # Create a figure to display results
5 fig, axes = plt.subplots(5, 10, figsize=(12, 6),
6                           subplot_kw={'xticks': [], 'yticks': []},
7                           gridspec_kw=dict(hspace=0.1, wspace=0.1))
8
9 # Display the first 50 test images with predictions
10 for i, ax in enumerate(axes.flat):

```



```

11     if i < 50: # Only fill the first 50 slots
12         ax.imshow(test_images[i], cmap='binary')
13
14         # Set text color based on whether prediction was correct
15         text_color = 'green' if y_pred[i] == y_test[i] else 'red'
16
17         # Display the predicted digit
18         ax.text(0.05, 0.05, str(y_pred[i]),
19               transform=ax.transAxes, color=text_color)
20
21 plt.suptitle('Digit Classification Results (green=correct, red=incorrect)')
22 plt.show()

```

Listing 26: Visualizing classification results

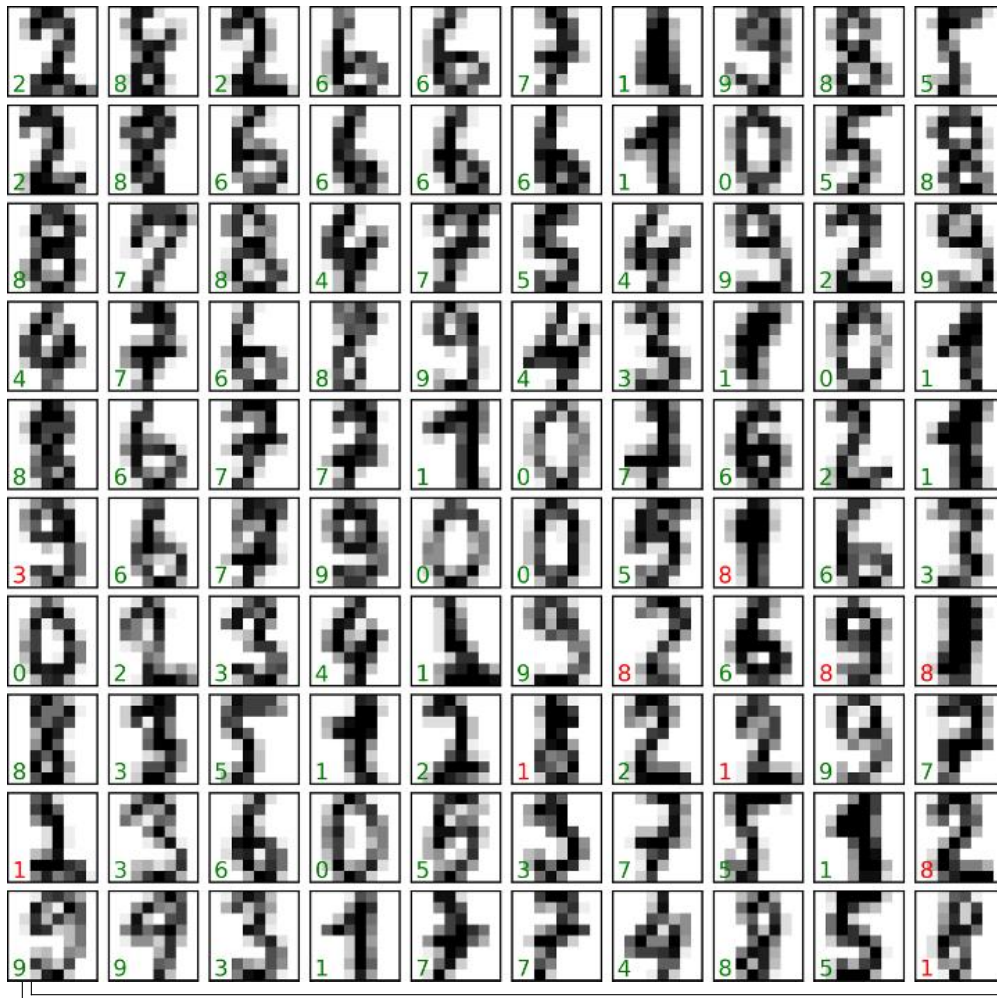


Figure 11: Test digits with their predicted labels (green = correct, red = incorrect).

### Why Do Models Make Mistakes?

Looking at the mistakes can be insightful:

- Some digits naturally look similar (e.g., 5 and 8, 1 and 7)
- Certain writing styles make recognition more difficult
- Our simple model doesn't capture all the nuances of handwriting

More sophisticated models like Support Vector Machines, Random Forests, or Convolutional Neural Networks could potentially achieve higher accuracy.

## 7 Step-by-Step Process: A Comprehensive Guide

To summarize and expand on the Scikit-Learn workflow, here's a comprehensive guide to approaching machine learning problems:

### 7.1 1. Understand the Problem

- Define what you're trying to predict or understand
- Determine if it's a supervised or unsupervised learning task
- For supervised learning, identify if it's classification or regression
- Consider how you'll evaluate success

### 7.2 2. Explore and Prepare the Data

- Examine data types, distributions, and relationships
- Handle missing values, outliers, and duplicates
- Encode categorical variables (one-hot encoding, label encoding)
- Scale numerical features if necessary
- Create new features through feature engineering
- Split data into training and testing sets

```
1 # Examine data
2 print(df.info())
3 print(df.describe())
4
5 # Handle missing values
6 df.fillna(df.mean(), inplace=True) # Fill with mean (numerical)
7 df['categorical_col'].fillna(df['categorical_col'].mode()[0], inplace=
   True) # Fill with mode
8
9 # Encode categorical variables
10 from sklearn.preprocessing import OneHotEncoder
11 encoder = OneHotEncoder(sparse=False, drop='first')
12 encoded_features = encoder.fit_transform(df[['categorical_col']])
13
14 # Scale numerical features
15 from sklearn.preprocessing import StandardScaler
16 scaler = StandardScaler()
17 scaled_features = scaler.fit_transform(df[numerical_columns])
18
19 # Split data
20 from sklearn.model_selection import train_test_split
21 X_train, X_test, y_train, y_test = train_test_split(
22     X, y, test_size=0.2, random_state=42, stratify=y) # stratify
   ensures class balance
```



### Warning

Common data preparation mistakes to avoid:

- **Data leakage:** Allowing information from the test set to influence training
- **Scaling after splitting:** Always fit scalers on training data only
- **Not handling categorical data:** Most algorithms require numerical input
- **Ignoring class imbalance:** Can lead to misleading accuracy metrics

## 7.3 3. Select a Model

- Consider the nature of your data and problem
- Start with simpler models before moving to complex ones
- Consider model interpretability vs. performance trade-offs
- Research which models perform well for similar problems

### Common Models by Task Type

- **Classification:**
  - Logistic Regression
  - Decision Trees
  - Random Forests
  - Support Vector Machines
  - Gradient Boosting
  - Neural Networks
- **Regression:**
  - Linear Regression
  - Ridge/Lasso Regression
  - Decision Trees
  - Random Forests
  - Gradient Boosting
  - Neural Networks
- **Clustering:**
  - K-means
  - DBSCAN
  - Hierarchical Clustering
  - Gaussian Mixture Models
- **Dimensionality Reduction:**

- PCA
- t-SNE
- UMAP
- Isomap

## 7.4 4. Train and Tune the Model

- Initialize the model with appropriate hyperparameters
- Fit the model to your training data
- Use cross-validation to evaluate performance
- Tune hyperparameters using grid search or random search

```

1 from sklearn.model_selection import cross_val_score, GridSearchCV
2 from sklearn.ensemble import RandomForestClassifier
3
4 # Initialize model
5 model = RandomForestClassifier(random_state=42)
6
7 # Cross-validation
8 cv_scores = cross_val_score(model, X_train, y_train, cv=5)
9 print(f"Cross-validation scores: {cv_scores}")
10 print(f"Mean CV score: {cv_scores.mean():.4f}")
11
12 # Hyperparameter tuning with GridSearchCV
13 param_grid = {
14     'n_estimators': [50, 100, 200],
15     'max_depth': [None, 10, 20, 30],
16     'min_samples_split': [2, 5, 10],
17     'min_samples_leaf': [1, 2, 4]
18 }
19
20 grid_search = GridSearchCV(
21     estimator=model,
22     param_grid=param_grid,
23     cv=5,
24     n_jobs=-1,
25     verbose=1,
26     scoring='accuracy'
27 )
28
29 grid_search.fit(X_train, y_train)
30
31 # Get best parameters and score
32 print(f"Best parameters: {grid_search.best_params_}")
33 print(f"Best cross-validation score: {grid_search.best_score_:.4f}")
34
35 # Train final model with best parameters
36 best_model = grid_search.best_estimator_

```

Listing 28: Cross-validation and hyperparameter tuning

## Understanding Cross-Validation

Cross-validation helps assess how a model will generalize to unseen data:

1. Data is divided into k subsets (folds)
2. For each fold, the model is trained on k-1 folds and tested on the remaining fold
3. Results are averaged across all k iterations

This provides a more robust estimate of model performance than a single train-test split.

## 7.5 5. Evaluate the Model

- Apply the model to your test set
- Calculate appropriate evaluation metrics
- Visualize results (confusion matrix, ROC curve, etc.)
- Compare against baseline models or business requirements

```
1 from sklearn.metrics import (accuracy_score, classification_report,
2                               confusion_matrix, roc_curve, auc)
3
4 # Make predictions
5 y_pred = best_model.predict(X_test)
6 y_pred_proba = best_model.predict_proba(X_test)[:, 1] # For binary
7               classification
8
9 # Basic metrics
10 accuracy = accuracy_score(y_test, y_pred)
11 print(f"Accuracy: {accuracy:.4f}")
12
13 # Detailed classification report
14 print("\nClassification Report:")
15 print(classification_report(y_test, y_pred))
16
17 # Confusion Matrix
18 cm = confusion_matrix(y_test, y_pred)
19 plt.figure(figsize=(8, 6))
20 sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
21 plt.xlabel('Predicted')
22 plt.ylabel('True')
23 plt.title('Confusion Matrix')
24 plt.show()
25
26 # ROC Curve (for binary classification)
27 fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
28 roc_auc = auc(fpr, tpr)
29
30 plt.figure(figsize=(8, 6))
31 plt.plot(fpr, tpr, color='darkorange', lw=2,
32          label=f'ROC curve (area = {roc_auc:.2f})')
33 plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
34 plt.xlim([0.0, 1.0])
35 plt.ylim([0.0, 1.05])
```

```

35 plt.xlabel('False Positive Rate')
36 plt.ylabel('True Positive Rate')
37 plt.title('Receiver Operating Characteristic')
38 plt.legend(loc="lower right")
39 plt.show()

```

Listing 29: Comprehensive model evaluation

## 7.6 6. Deploy and Monitor the Model

- Save the trained model for future use
- Implement the model in production systems
- Monitor performance over time
- Retrain as needed with new data

```

1 import joblib
2
3 # Save the model
4 joblib.dump(best_model, 'best_model.pkl')
5
6 # Load the model later
7 loaded_model = joblib.load('best_model.pkl')
8
9 # Use the loaded model
10 new_predictions = loaded_model.predict(new_data)

```

Listing 30: Saving and loading models

### Model Deployment Best Practices

- Save all preprocessing steps (scalers, encoders) along with the model
- Create a pipeline that combines preprocessing and the model
- Implement monitoring for concept drift (when data distribution changes)
- Set up automated retraining when performance drops
- Document the entire workflow for future maintainers

## 8 Summary

In this chapter, we covered the essential features of the Scikit-Learn API and how to use it for machine learning tasks:

- **Data Representation:** Features matrix (X) and target vector (y)
- **Estimator API:** A consistent pattern of importing, instantiating, fitting, and applying models
- **Supervised Learning:** Examples with linear regression and classification
- **Unsupervised Learning:** Examples with dimensionality reduction and clustering

- **Application:** A more complex example with handwritten digit recognition
- **Comprehensive Process:** A detailed workflow for addressing machine learning problems

The power of Scikit-Learn lies in its consistent API—once you understand the basic pattern, you can apply it to a wide range of machine learning tasks with minimal changes to your code structure.

In the next chapters, you'll explore specific algorithms in more depth and learn techniques for selecting and validating your models to ensure they generalize well to new data.

### Next Steps

To continue your Scikit-Learn journey:

- Experiment with different algorithms for your specific problems
- Learn about feature engineering to improve model performance
- Explore model evaluation techniques in greater depth
- Study hyperparameter tuning to optimize your models
- Try more advanced techniques like ensemble methods