The Machine Learning Workflow: Takeaways

Syntax

- Loading the Breast Cancer Wisconsin (Diagnostic) Dataset in scikit-learn into a Pandas DataFrame:

```
cancer_data = load_breast_cancer(as_frame=True)
cancer_df = cancer_data.data
cancer_df['target'] = cancer_data.target
```

- Preparing and splitting the dataset into training and test datasets:

```
from sklearn.model_selection import train_test_split
X = cancer_df.drop(["target"], axis=1)
y = cancer_df["target"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.15, random_state=417)
```

- Building and training a LinearSVC model in scikit-learn:

```
model = LinearSVC(penalty="l2", loss="squared_hinge", C=10, random_state=417)
model.fit(X_train, y_train)
```

- Calculating the accuracy of a model in scikit-learn:

```
model.score(X_test, y_test)
```

Concepts

- **Machine Learning** enables us to build mathematical **models** that can identify and learn patterns from given data on their own and make predictions on unseen data.
- The **machine learning workflow** contains the following series of steps:
    - Data Collection.
    - Data Exploration and Wrangling.
    - Data Preparation.
    - Building and training a model.
    - Evaluating the model performance.
    - Fine-tuning the model.
    - Evaluating the model performance.
- For tabular data:
    - A column is called a **feature**. It describes or is a property of our data.
    - The feature that we want our model to predict is called the **target variable**.
    - A row is called an **observation** or **feature vector**.
- In **supervised machine learning**, a machine learning model learns from, or is trained on, data that has labels or classes for every observation.
    - The labels can be continuous or categorical values.

- For **classification** tasks, the model, also called a **classifier**, learns to predict categorical labels.
    - If there are only two labels, it's called **binary classification**.
    - If there are more than two labels, it's called **multi-class classification**.
- Data used to train a model is called **training data**, a **training set** or a **training dataset**.
- Data used to test a model is called **test data**,a **test set** or a **test dataset**.
    - The test set can be obtained by splitting the original dataset into a training and test set.
    - The test set's size is usually about `15` to `20` % of the original dataset.
- A classifier's performance can be evaluated by calculating its **accuracy**. The accuracy of a model can be calculated by comparing those predictions to the actual labels.
- Every model has its own set of parameters that can be tuned as an attempt to improve the model's performance.

Resources

- Breast Cancer Wisconsin (Diagnostic) Dataset
- pandas' shape() function
- pandas' isna() function
- scikit-learn's train_test_split() method
- scikit-learn's Linear Support Vector Classification model
- scikit-learn's score() function