

*Worksheet 7 - Oct 24, 2025*

**Intro to Machine Learning**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dataset:** [spotify\_mini.csv](https://github.com/Sammyjoon/Intro-to-Data-Science/tree/main/Week%206)

### 🧠 ****Goal****

Today, you’ll explore how computers learn from data!  
We’ll use a small Spotify dataset to understand:

* What **features** and **labels** are
* The difference between **training** and **testing** data

🧩 Step 1: Load & Explore the Data

1️. Download **spotify\_mini.csv** from GitHub (Week 6) and upload it to your Colab sidebar (Files ➜ Upload).  
2️. Run this code 👇:

A close-up of a computer code

Description automatically generated

1. 📝 **Question 1:**  
   What columns do you see in the dataset? List three columns you find interesting.

|  |  |  |
| --- | --- | --- |
| **column 1** | **column 2** | **column 3** |
|  |  |  |
|  |  |  |

### 📊 **Step 2: Choose a Feature and Label**

We’ll try to predict how many **streams** a song gets based on one feature (for example, skip\_rate).

A white background with black text

Description automatically generated

📝 **Question 2:**

* What is the **feature** in this example?
* What is the **label**?

|  |  |
| --- | --- |
| **Feature** | **Label** |
|  |  |
|  |  |

✂️ **Step 3: Split the Data into Training and Testing Sets**

We want to see if the computer can learn patterns — so we’ll train it on part of the data and test it on the rest.

A white background with black and pink text

Description automatically generated

📝 **Question 3:**

* How much of the data is used for training?
* How much is used for testing?
* Why do we split the data?

|  |  |  |
| --- | --- | --- |
| **Training %** | **Testing %** | **Why do we split?** |
|  |  |  |
|  |  |  |

👀 **Step 4: Check Your Split**

Run this to see how many examples are in each part:

A close-up of text

Description automatically generated

1. 📝 **Question 4:**  
   Write down how many rows your training and testing sets have:

|  |  |
| --- | --- |
| **Data Type** | **Number of Rows** |
| Training |  |
| Testing |  |

### 💬 ****Step 5: Reflection****

1. Why shouldn’t we train and test on the same data?
2. What might go wrong if our data only included songs from one genre?

### ⭐ ****Bonus Challenge****

Change test\_size to 0.3 or 0.5.  
How do the training and testing sizes change?