# 5 Free Datasets to Start Your Machine Learning Projects

Use regression, classification, time series, image, and text dataset to work on portfolio projects.



Image created with ChatGPT and Canva

After completing a course and mastering the essentials of machine learning, it is time to start building machine learning models using real-world datasets. While accessing most real-world datasets can be expensive, platforms like Kaggle offer access to the latest datasets for free on a wide variety of machine learning tasks.

In this blog, we will explore 5 free datasets that you can use to build a strong machine learning portfolio. By using these datasets, we will be able to build regression, classification, time series, computer vision, and natural language processing models, providing a comprehensive foundation for your machine learning journey.

## 1. Boston House Prices

[Link to Dataset](https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data)

The Boston House Prices dataset is a classic dataset widely used for regression tasks. It is ideal for practicing a variety of regression techniques, such as linear regression, decision trees, and more advanced methods. By cleaning and preprocessing the data, and fitting it to your models, you can predict house prices based on various features like the number of rooms, crime rate, age, and tax rate. This dataset provides a comprehensive platform to enhance your skills in data manipulation and model building.

## 2. Stroke Prediction Dataset

[Link to Dataset](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset)

The Stroke Prediction dataset is a valuable tool for predicting whether a patient is likely to suffer a stroke based on various input features. These features include gender, age, the presence of diseases like hypertension and heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status. The dataset is ideal for building classification models such as logistic regression, random forests, or neural networks.

## 3. Netflix Stock Price Prediction

[Link to Dataset](https://www.kaggle.com/datasets/jainilcoder/netflix-stock-price-prediction)

The Netflix Stock Price Prediction dataset is perfect for time series analysis. It provides historical stock price data for Netflix, including open, high, low, close prices, and volume. This dataset is suitable for building models to predict future stock prices using techniques like ARIMA, LSTM, or other time series forecasting models. Financial datasets like this one are crucial for those interested in working in the financial field and building the algorithms for trading.

## 4. ImageNet

[Link to Dataset](https://image-net.org/download.php)

ImageNet is one of the largest and most well-known datasets for computer vision tasks. It contains millions of images with labels spanning thousands of categories. This dataset is essential for training deep learning models such as Convolutional Neural Networks (CNNs) for image classification, object detection, and segmentation. ImageNet is a gold standard in the field of computer vision and is used to benchmark the performance of new algorithms.

## 5. Yelp Dataset

[Link to Dataset](https://www.yelp.com/dataset)

The Yelp Dataset is a comprehensive dataset for natural language processing (NLP) tasks. It includes information on businesses, reviews, and user data from Yelp. This dataset is ideal for sentiment analysis, recommendation systems, and various text classification tasks. By using this dataset, you can practice building models that understand human-like text, which is a crucial skill in this day and age when everyone is obsessed with AI and large language models.

## Conclusion

It is crucial to remember that building a strong machine learning portfolio requires practical experience with real-world datasets. The five datasets discussed in this blog cover a wide variety of machine learning tasks, including regression, classification, time series analysis, computer vision, and natural language processing. By working with these datasets, you can develop a comprehensive skill set that will establish a solid foundation for your machine learning career.