<https://machinelearningmastery.com/standard-machine-learning-datasets/>

The key to getting good at applied machine learning is practicing on lots of different datasets.

This is because each problem is different, requiring subtly different data preparation and modeling methods.

In this post, you will discover 10 top standard machine learning datasets that you can use for practice.

Let’s dive in

**Overview**

**A structured Approach**

Each dataset is summarized in a consistent way. This makes them easy to compare and navigate for you to practice a specific data preparation technique or modeling method.

The aspects that you need to know about each dataset are:

1. **Name**: How to refer to the dataset.
2. **Problem Type**: Whether the problem is regression or classification.
3. **Inputs and Outputs**: The numbers and known names of input and output features.
4. **Performance**: Baseline performance for comparison using the Zero Rule algorithm, as well as best known performance (if known).
5. **Sample**: A snapshot of the first 5 rows of raw data.
6. **Links**: Where you can download the dataset and learn more.

**Standard Datasets**

Below is a list of the 10 datasets we’ll cover.

Each dataset is small enough to fit into memory and review in a spreadsheet. All datasets are comprised of tabular data and no (explicitly) missing values.

1. Swedish Auto Insurance Dataset.
2. Wine Quality Dataset.
3. Pima Indians Diabetes Dataset.
4. Sonar Dataset.
5. Banknote Dataset.
6. Iris Flowers Dataset.
7. Abalone Dataset.
8. Ionosphere Dataset.
9. Wheat Seeds Dataset.
10. Boston House Price Dataset.

**Things to Complete:**

Exploratory Data Analysis: Due 23rd July

Decision Tree: Due 30th July

Regression: Due 6th August

Final project: Due 13th August.

**Banknote Dataset**

The Banknote Dataset involves predicting whether a given banknote is authentic given a number of measures taken from a photograph.

It is a binary (2-class) classification problem. The number of observations for each class is not balanced. There are 1,372 observations with 4 input variables and 1 output variable. The variable names are as follows:

1. Variance of Wavelet Transformed image (continuous).
2. Skewness of Wavelet Transformed image (continuous).
3. Kurtosis of Wavelet Transformed image (continuous).
4. Entropy of image (continuous).
5. Class (0 for authentic, 1 for inauthentic).

The baseline performance of predicting the most prevalent class is a classification accuracy of approximately 50%.

A sample of the first 5 rows is listed below.

3.6216,8.6661,-2.8073,-0.44699,0

4.5459,8.1674,-2.4586,-1.4621,0

3.866,-2.6383,1.9242,0.10645,0

3.4566,9.5228,-4.0112,-3.5944,0

0.32924,-4.4552,4.5718,-0.9888,0

4.3684,9.6718,-3.9606,-3.1625,0

Download Dataset: <http://archive.ics.uci.edu/ml/machine-learning-databases/00267/data_banknote_authentication.txt>

More Information: <http://archive.ics.uci.edu/ml/datasets/banknote+authentication>

Please follow the following format for the next presentation:

1. Project description slide
2. EDA
3. Rules