# Classification – Practice Areas

**Practice Problems:**

**Predict Students' Dropout and Academic Success**

A dataset created from a higher education institution (acquired from several disjoint databases) related to students enrolled in different undergraduate degrees, such as agronomy, design, education, nursing, journalism, management, social service, and technologies. The dataset includes information known at the time of student enrollment (academic path, demographics, and social-economic factors) and the students' academic performance at the end of the first and second semesters. The data is used to build classification models to predict students' dropout and academic sucess. The problem is formulated as a three category classification task, in which there is a strong imbalance towards one of the classes.

Classification

Tabular

4.42K Instances

36 Features

**Product Classification and Clustering**

This dataset was collected from PriceRunner, a popular product comparison platform. It includes 35311 product offers from 10 categories, provided by 306 different merchants. This dataset offers an ideal ground for evaluating classification, clustering, and entity matching algorithms. Although it contains product-related data, it can still be applied to any problem involving text/short-text mining.

Classification, Clustering, Other

Tabular, Text

35.31K Instances

7 Features

## Multivariate Gait Data

Bilateral (left, right) joint angle (ankle, knee, hip) times series data collected from 10 healthy subjects under 3 walking conditions (unbraced, knee braced, ankle braced). For each condition, each subject's data consists of 10 consecutive gait cycles.

Classification, Regression, Clustering

Sequential, Multivariate, Time-Series

181.8K Instances

7 Features

## RealWaste

An image classification dataset of waste items across 9 major material types, collected within an authentic landfill environment.

Classification

Image

4.75K Instances

## Recipe Reviews and User Feedback

The "Recipe Reviews and User Feedback Dataset" is a comprehensive repository of data encompassing various aspects of recipe reviews and user interactions. It includes essential information such as the recipe name, its ranking on the top 100 recipes list, a unique recipe code, and user details like user ID, user name, and an internal user reputation score. Each review comment is uniquely identified with a comment ID and comes with additional attributes, including the creation timestamp, reply count, and the number of up-votes and down-votes received. Users' sentiment towards recipes is quantified on a 1 to 5 star rating scale, with a score of 0 denoting an absence of rating. This dataset is a valuable resource for researchers and data scientists, facilitating endeavors in sentiment analysis, user behavior analysis, recipe recommendation systems, and more. It offers a window into the dynamics of recipe reviews and user feedback within the culinary website domain.

Classification, Other

Tabular, Other

18.18K Instances

15 Features

## Bengali Hate Speech Detection Dataset

The dataset can be used for hate speech detection in Bengali social media texts. The dataset is categorized into political, personal, geopolitical, religious, and gender abusive hates that are either directed or generalized towards a specific person, entity, or group. The data and lexicons contain content that is racist, sexist, homophobic, and offensive in many different ways. The dataset is collected and subsequently annotated only for research-related purposes. Besides, authors don't take any liability if some statements contain very offensive and hateful statements that are either directed towards a specific person or entity or generalized towards a group. Therefore, please use it at your risk.

Classification

Text

4.5K Instances

## Sundanese Twitter Dataset

This dataset contains tweet of the second-largest local language in Indonesia and is used for emotion classification.

Classification

Tabular

2.51K Instances

1 Features

## Inflation Research Abstracts Classification

This data set contains scientific papers abstracts from economics inflation. The task is to classify them according to their machine learning methodologies inclusion.

Classification

Text

1.14K Instances

**Biological and Medical Datasets**

Iris Dataset

- The Iris dataset is a classic dataset in the field of machine learning, consisting of 150 observations of iris flowers.
- Each observation has four features (sepal length, sepal width, petal length, petal width) and belongs to one of three species: Setosa, Versicolour, or Virginica. It is commonly used for classification tasks and visualizations.

Breast Cancer Wisconsin Dataset

- Breast Cancer Wisconsin Dataset contains features computed from breast cancer biopsy images, aiming to predict whether a tumor is benign or malignant. It includes 569 instances with 30 features such as radius, texture, perimeter, and area of the nuclei.
- It is widely used in the medical field for diagnostic purposes.

Heart Disease Dataset

- The Heart Disease dataset contains various patient attributes to predict the presence of heart disease. It includes features like age, sex, chest pain type, resting blood pressure, and cholesterol levels, with a total of 303 instances.
- This dataset is essential for developing models to diagnose cardiovascular conditions.

**Finance and Socio-economic Datasets**

Titanic Dataset

- The Titanic dataset provides information about the passengers aboard the Titanic, used to predict survival rates. It includes features such as passenger class, age, gender, ticket fare, and whether they had family on board.
- This dataset is popular for binary classification and feature engineering tasks.

Adult Census Income Dataset

- Also known as the "Census Income" dataset, it contains demographic information from the 1994 Census database to predict whether an individual earns more than $50,000 a year.
- It has 48,842 instances with 14 attributes like age, work class, education, marital status, and occupation.
- It can be obtained from official website.

**Image Classification Datasets**

MNIST Dataset:

- The MNIST dataset is a collection of 70,000 handwritten digit images (0-9) used for image classification. Each image is 28x28 pixels, with 60,000 images for training and 10,000 for testing.
- It is a fundamental dataset for beginners in computer vision and deep learning.

Digits Dataset:

- Similar to MNIST, the Digits dataset contains images of handwritten digits (0-9) from the **scikit-learn** library.
- It includes 1,797 grayscale images of 8x8 pixels, used for classification tasks and algorithm comparisons in image recognition.

Fashion MNIST Dataset:

- Fashion MNIST is a dataset of 70,000 grayscale images of 10 fashion categories (e.g., T-shirts, trousers, bags, shoes).
- Each image is 28x28 pixels, intended as a more challenging drop-in replacement for the original MNIST dataset, promoting more advanced research in computer vision.

Chemical Analysis and Manufacturing Dataset

Wine Dataset

- The Wine dataset consists of 178 instances of Italian wines, classified into three types.
- Each instance is described by 13 chemical properties like alcohol content, malic acid, ash, and color intensity. It is widely used for classification and clustering in chemical and quality control analysis.

Text and Natural Language Processing Dataset

Spam Email Dataset

- The Spam Email dataset contains email messages labeled as spam or non-spam, used for spam detection. It includes features derived from the email content, such as word frequencies and the presence of certain keywords.
- This dataset is crucial for developing and testing email filtering algorithms.

Other dataset for practice for linear regression:

https://www.kaggle.com/datasets?tags=13302-Classification

(Unlimited practice for linear regression)

All the best.

Let's refine and refine.

Thanks.