

Package dsf

2025-08-30

Version 0.0.7

Description The **dsf** package provides a collection of datasets used in the book **Data Science Foundations and Machine Learning with Python**. It is designed to make data science techniques accessible to individuals with minimal coding experience.

URL <https://github.com/vanraak/dsf>

Depends Python (≥ 3.8) and Pandas (>2.0)

License GPL (≥ 2)

Repository Pypi

Authors Jeroen van Raak and Reza Mohammadi

Maintainer Jeroen van Raak, j.j.f.vanraak@uva.nl

Installation

```
pip install dsf
```

Usage

```
import dsf
df=dsf.load('<dataset>')
```

Replace with the name of the dataset, such as 'bank', 'house', or 'churn'.

Example

```
df=dsf.load('bank') # Load the bank dataset.
```

Datasets

adult	3
advertising	5
bank	6
caravan	8
cereal	9
churn	11
churncredit	13
churrtel	15
corona	17
diamonds	18
fertilizer	19
house	20
houseprice	21
insurance	22
marketing	23
redwines	24
risk	26
whitewines	27

adult Adult dataset

Description

The adult dataset was collected from the US Census Bureau and the primary task is to predict whether a given adult makes more than \$50K a year based attributes such as education, hours of work per week, etc. the target feature is *income*, a factor with levels “<=50K” and “>50K”, and the remaining 14 variables are predictors.

Usage

```
df=dsf.load('adult')
```

Format

The adult dataset, as a data frame, contains 48598 rows and 15 columns (variables/features).

The 15 variables are:

- age: age in years.
- workclass: a factor with 6 levels.
- demogweight: the demographics to describe a person.
- education: a factor with 16 levels.
- education_num: number of years of education.
- marital_status: a factor with 5 levels.
- occupation: a factor with 15 levels.
- relationship: a factor with 6 levels.
- race: a factor with 5 levels.
- gender: a factor with levels “Female”, “Male”.
- capital_gain: capital gains.
- capital_loss: capital losses.
- hours_per_week: number of hours of work per week.
- native_country: a factor with 42 levels.
- income: yearly income as a factor with levels “<=50K” and “>50K”.

Details

This dataset can be downloaded from the UCI machine learning repository:

<http://www.cs.toronto.edu/~dave/data/adult/desc.html>

A detailed description of the dataset can be found in the UCI documentation at:

<http://www.cs.toronto.edu/~dave/data/adult/adultDetail.html>

References

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. *Kdd*.

advertising Advertising dataset

Description

The dataset is from an anonymous organisation's social media ad campaign. The advertising dataset contains 11 features and 1143 records.

Usage

```
df=dsf.load('advertising')
```

Format

The advertising dataset, as a data frame, contains 1143 rows and 11 columns (variables/features).

The 11 variables are:

- `ad_id`: an unique ID for each ad.
- `xyz_campaign_id`: an ID associated with each ad campaign of XYZ company.
- `fb_campaign_id`: an ID associated with how Facebook tracks each campaign.
- `age`: age of the person to whom the ad is shown.
- `gender`: gender of the person to whom the ad is shown.
- `interest`: a code specifying the category to which the person's interest belongs (interests are as mentioned in the person's Facebook public profile).
- `impressions`: the number of times the ad was shown.
- `clicks`: number of clicks on for that ad.
- `spend`: amount paid by company xyz to Facebook, to show that ad.
- `conversion`: total number of people who enquired about the product after seeing the ad.
- `approved`: total number of people who bought the product after seeing the ad.

Details

A detailed description of the dataset can be found at:

<https://www.kaggle.com/loveall/clicks-conversion-tracking>

bank Bank Marketing dataset

Description

The data is related to direct marketing campaigns of a Portuguese banking institution. the marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. the classification goal is to predict if the client will subscribe a term deposit (variable deposit).

Usage

```
df=dsf.load('bank')
```

Format

The bank dataset, as a data frame, contains 4521 rows (customers) and 17 columns (variables/features).

The 17 variables are:

Bank client data:

- age: numeric.
- job: type of job; categorical: “admin.”, “unknown”, “unemployed”, “management”, “housemaid”, “entrepreneur”, “student”, “blue-collar,”self-employed”, “retired”, “technician”, “services”.
- marital: marital status; categorical: “married”, “divorced”, “single”; note: “divorced” means divorced or widowed.
- education: categorical: “secondary”, “primary”, “tertiary”, “unknown”.
- default: has credit in default?; binary: “yes”, “no”.
- balance: average yearly balance, in euros; numeric.
- housing: has housing loan? binary: “yes”, “no”.
- loan: has personal loan? binary: “yes”, “no”.

Related with the last contact of the current campaign:

- contact: contact: contact communication type; categorical: “unknown”, “telephone”, “cellular”.
- day: last contact day of the month; numeric.
- month: last contact month of year; categorical: “jan”, “feb”, “mar”, ..., “nov”, “dec”.
- duration: last contact duration, in seconds; numeric.

Other attributes:

- campaign: number of contacts performed during this campaign and for this client; numeric, includes last contact.
- pdays: number of days that passed by after the client was last contacted from a previous campaign; numeric, -1 means client was not previously contacted.
- previous: number of contacts performed before this campaign and for this client; numeric.
- poutcome: outcome of the previous marketing campaign; categorical: “success”, “failure”, “unknown”, “other”.

Target variable:

- deposit: Indicator of whether the client subscribed a term deposit; binary: “yes” or “no”.

Details

This dataset can be downloaded from the UCI machine learning repository:

<http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

References

Moro, S., Laureano, R. and Cortez, P. (2011) Using Data Mining for Bank Direct Marketing: An

Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference.

caravan Caravan Insurance dataset

Description

The dataset contains 5822 customer records from an insurance company, each described by 86 variables. These include 43 sociodemographic features based on zip codes and 43 indicators of product ownership. The final variable, Purchase, indicates whether a customer bought a caravan insurance policy. Collected for the CoIL 2000 Challenge, the data was designed to address the question: Can you predict who would be interested in buying a caravan insurance policy and explain why?

Usage

```
df=dsf.load('bank')
```

Format

The caravan dataset contains 5822 rows (customers) and 86 columns (variables/features).

Further variable details are available at <http://www.liacs.nl/~putten/library/cc2000/data.html>.

Details

The data was supplied by Sentient Machine Research: <https://www.smr.nl>

References

P. van der Putten and M. van Someren (eds) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000. <http://www.liacs.nl/~putten/library/cc2000>.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with applications in R, <https://www.statlearning.com>, Springer-Verlag.

cereal Cereal dataset

Description

This dataset contains nutrition information for 77 breakfast cereals and includes 16 variables. The “rating” column is our target as a rating of the cereals.

Usage

```
df=dsf.load('cereal')
```

Format

The cereal dataset, as a data frame, contains 77 rows (breakfast cereals) and 16 columns (variables/features).

The 16 variables are:

- name: Name of cereal.
- manuf: Manufacturer of cereal:
 - A: American Home Food Products;
 - G: General Mills;
 - K: Kelloggs;
 - N: Nabisco;
 - P: Post;
 - Q: Quaker Oats;
 - R: Ralston Purina;
- type: cold or hot.
- calories: calories per serving.
- protein: grams of protein.
- fat: grams of fat.
- sodium: milligrams of sodium.
- fiber: grams of dietary fiber.
- carbo: grams of complex carbohydrates.
- sugars: grams of sugars.
- potass: milligrams of potassium.
- vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended.
- shelf: display shelf (1, 2, or 3, counting from the floor).
- weight: weight in ounces of one serving.
- cups: number of cups in one serving.
- rating: a rating of the cereals (Possibly from Consumer Reports?).

Details

The original source can be found at:

<https://perso.telecom-paristech.fr/eagan/class/igr204/datasets>

churn Churn dataset

Description

This dataset comes from IBM Sample Data Sets. Customer *churn* occurs when customers stop doing business with a company, also known as customer attrition. the dataset contains 5000 rows (customers) and 20 columns (features). the “Churn” column is our target which indicate whether customer churned (left the company) or not.

Usage

```
df=dsf.load('churn')
```

Format

The churn dataset, as a data frame, contains 5000 rows (customers) and 20 columns (variables/features).

The 20 variables are:

- state: Categorical, for the 51 states and the District of Columbia.
- area_code: Categorical.
- account_length: count, how long account has been active.
- voice_plan: Categorical, yes or no, voice mail plan.
- voice_messages: Count, number of voice mail messages.
- intl_plan: Categorical, yes or no, international plan.
- intl_mins: Continuous, minutes customer used service to make international calls.
- intl_calls: Count, total number of international calls.
- intl_charge: Continuous, total international charge.
- day_mins: Continuous, minutes customer used service during the day.
- day_calls: Count, total number of calls during the day.
- day_charge: Continuous, total charge during the day.
- eve_mins: Continuous, minutes customer used service during the evening.
- eve_calls: Count, total number of calls during the evening.
- eve_charge: Continuous, total charge during the evening.
- night_mins: Continuous, minutes customer used service during the night.
- night_calls: Count, total number of calls during the night.
- night_charge: Continuous, total charge during the night.
- customer_calls: Count, number of calls to customer service.
- churn: Categorical, yes or no. Indicator of whether the customer has left the company (yes or no).

References

Larose, D. T. and Larose, C. D. (2014). Discovering knowledge in data: an introduction to data mining. *John Wiley & Sons*.

churncredit Churn dataset for Credit Card Customers

Description

Customer *churn* occurs when customers stop doing business with a company, also known as customer attrition. the dataset contains 10127 rows (customers) and 21 columns (features). The “churn” column is our target which indicate whether customer churned (left the company) or not.

Usage

```
df=dsf.load('churncredit')
```

Format

The churncredit dataset, as a data frame, contains 10127 rows (customers) and 21 columns (variables/features).

The 21 variables are:

- customer_id: Customer ID.
- gender: Whether the customer is a male or a female.
- age: Customer's Age in Years.
- educaton: Educational Qualification of the account holder (example: high school, college graduate, etc.)
- marital_status: Married, Single, Divorced, Unknown
- income: Annual Income (in Dollar). Category of the account holder (< \$40K, \$40K - 60K, \$60K - \$80K, \$80K-\$120K, > \$120K).
- dependent_counts: Number of dependent counts.
- card_category: Type of Card (Blue, Silver, Gold, Platinum).
- months_on_book: Period of relationship with bank.
- relationship_count: Total number of products held by the customer.
- months_inactive: Number of months inactive in the last 12 months.
- contacts_count_12: Number of Contacts in the last 12 months.
- credit_limit: Credit Limit on the Credit Card.
- revolving_balance: Total Revolving Balance on the Credit Card.
- open_to_buy: Open to Buy Credit Line (Average of last 12 months).
- transaction_amount_Q4_Q1: Change in Transaction Amount (Q4 over Q1).
- transaction_amount_12: Total Transaction Amount (Last 12 months).
- transaction_count: Total Transaction Count (Last 12 months).
- transaction_change: Change in Transaction Count (Q4 over Q1).
- utilization_ratio: Average Card Utilization Ratio.
- churn: Whether the customer churned or not (yes or no).

Details

For more information related to the dataset see:

<https://www.kaggle.com/sakshigoyal7/credit-card-customers>

churntel Telco Customer Churn dataset

churnTel dataset

Description

Customer *churn* occurs when customers stop doing business with a company, also known as customer attrition. the dataset contains 7043 rows (customers) and 21 columns (features). The “Churn” column is our target which indicate whether customer churned (left the company) or not.

Usage

```
df=dsf.load('churntel')
```

Format

The churnTel dataset, as a data frame, contains 7043 rows (customers) and 21 columns (variables/features). the 21 variables are:

- `customer_id`: Customer ID.
- `gender`: Whether the customer is a male or a female.
- `senior_citizen`: Whether the customer is a senior citizen or not (1, 0).
- `partner`: Whether the customer has a partner or not (yes, no).
- `dependent`: Whether the customer has dependents or not (yes, no).
- `tenure`: Number of months the customer has stayed with the company.
- `phone_service`: Whether the customer has a phone service or not (yes, no).
- `multiple_lines`: Whether the customer has multiple lines or not (yes, no, no phone service).
- `internet_service`: Customer’s internet service provider (DSL, fiber optic, no).
- `online_security`: Whether the customer has online security or not (yes, no, no internet service).
- `online_backup`: Whether the customer has online backup or not (yes, no, no internet service).
- `device_protection`: Whether the customer has device protection or not (yes, no, no internet service).
- `tech_support`: Whether the customer has tech support or not (yes, no, no internet service).
- `streaming_TV`: Whether the customer has streaming TV or not (yes, no, no internet service).
- `streaming_movie`: Whether the customer has streaming movies or not (yes, no, no internet service).
- `contract`: the contract term of the customer (month to month, 1 year, 2 year).

- `paperless_bill`: Whether the customer has paperless billing or not (yes, no).
- `payment_method`: the customer's payment method (electronic check, mail check, bank transfer, credit card).
- `monthly_charge`: the amount charged to the customer monthly.
- `total_charges`: the total amount charged to the customer.
- `churn`: Whether the customer churned or not (yes or no).

Details

For more information related to the dataset see: <https://www.kaggle.com/blastchar/telco-customer-churn>

corona COVID-19 Coronavirus dataset

Description

COVID-19 Coronavirus data - daily (up to 14 December 2020).

Usage

```
df=dsf.load('corona')
```

Format

The corona dataset, as a data frame, contains 61900 rows and 12 columns (variables/features).

Details

This dataset can be downloaded from European data repository:

<https://data.europa.eu/data/datasets/covid-19-coronavirus-data-daily-up-to-14-december-2020>

diamonds Diamonds dataset

Description

The diamonds dataset is a comprehensive collection of data about diamonds, widely used in data science to practice visualization, statistical modeling, and machine learning. It includes detailed characteristics of individual diamonds and their corresponding prices. The dataset contains detailed information on over 50,000 diamonds, including both physical characteristics and pricing.

Usage

```
df=dsf.load('diamonds')
```

Format

The 10 variables are:

- carat: weight of the diamond (ranging from 0.2 to 5.01),
- cut: quality of the cut (Fair, Good, Very Good, Premium, Ideal),
- color: color grade, from D (most colorless) to J (least colorless),
- clarity: clarity grade, from I1 (least clear) to IF (flawless),
- depth: total depth percentage calculated as: $2 * z / (x + y)$,
- table: width of the top facet relative to the widest point (43-95%),
- x: length in mm
- y: width in mm
- z: depth in mm
- price: price in US dollars (ranging from \$326 to \$18,823).

Details

For more information related to the dataset see:

<https://search.r-project.org/CRAN/refmans/nodbi/html/diamonds.html>

fertilizer Fertilizer dataset

Description

The fertilizer dataset contains 4 features and 96 records. Results from an experiment to compare yields of a crop obtained under three different fertilizers. The target feature is *yield*.

Usage

```
df=dsf.load('fertilizer')
```

house House dataset

Description

The house dataset contains 6 features and 414 records. The target feature is *unit_price* and the remaining 5 variables are predictors.

Usage

```
df=dsf.load('house')
```

Format

The house dataset, as a data frame, contains 414 rows and 6 columns (variables/features). The 6 variables are:

- house_age: house age (numeric, in year).
- distance_to_mrt: distance to the nearest MRT station (numeric).
- stores_number: number of convenience stores (numeric).
- latitude: latitude (numeric).
- longitude: longitude (numeric).
- unit_price: house price of unit area (numeric).

Details

A detailed description of the dataset can be found at:

<https://www.kaggle.com/quantbruce/real-estate-price-prediction>

houseprice HousePrice dataset

Description

This dataset contains 1460 rows and 81 columns (features). the “SalePrice” column is the target.

Usage

```
df=dsf.load('houseprice')
```

Format

The housePrice dataset, as a data frame, contains 1460 rows and 81 columns (variables/features).

Details

For more information related to the dataset see:

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

insurance Insurance dataset

Description

The insurance dataset contains 7 features and 1338 records. the target feature is *charge* and the remaining 6 variables are predictors.

Usage

```
df=dsf.load('insurance')
```

Format

The insurance dataset, as a data frame, contains 1338 rows (customers) and 7 columns (variables/features).

The 7 variables are:

- age: age of primary beneficiary.
- bmi: body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9.
- children: Number of children covered by health insurance / Number of dependents.
- smoker: Smoking as a factor with 2 levels, yes, no.
- gender: insurance contractor gender, female, male.
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charge: individual medical costs billed by health insurance.

Details

A detailed description of the dataset can be found at:

<https://www.kaggle.com/mirichoi0218/insurance>

References

Brett Lantz (2019). Machine Learning with R: Expert techniques for predictive modeling. *Packt Publishing Ltd*.

marketing Marketing dataset

Description

The marketing dataset contains 8 features and 40 records as 40 days that report how much we spent, how many clicks, impressions and transactions we got, whether or not a display campaign was running, as well as our revenue, click-through-rate and conversion rate. the target feature is revenue and the remaining 7 variables are predictors.

Usage

```
df=dsf.load('marketing')
```

Format

The marketing dataset, as a data frame, contains 40 rows and 8 columns (variables/features).

The 8 variables are:

- spend: daily spend of money on PPC (pay-per-click).
- clicks: number of clicks on for that ad.
- impressions: amount of impressions per day.
- display: whether or not a display campaign was running.
- transactions: number of transactions per day.
- click_rate: click-through-rate.
- conversion.rate: conversion rate.
- revenue: daily revenue.

Details

A detailed description of the dataset can be found at:

<https://github.com/chrisBow/marketing-regression-part-one>

redwines Red Wines dataset

Description

The redWines datasets are related to red variants of the Portuguese “Vinho Verde” wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The dataset can be viewed as classification or regression tasks. the classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Usage

```
df=dsf.load('redwines')
```

Format

The redWines dataset, as a data frame, contains 1599 rows and 12 columns (variables/features).

The 12 variables are:

Input variables (based on physicochemical tests):

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- pH
- sulphates
- alcohol

Output variable (based on sensory data)

- quality: score between 0 and 10.

Details

This dataset can be downloaded from the UCI machine learning repository:

<https://archive.ics.uci.edu/dataset/186/wine+quality>

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4), 547-553.

risk Risk dataset

Description

The *risk* dataset containing 6 features and 246 records. the target feature is *risk*, a factor with levels “good risk” and “bad risk” along with 5 predictors.

Usage

```
df=dsf.load('risk')
```

Format

The risk dataset, as a data frame, contains 246 rows (customers) and 6 columns (variables/features).

The 6 variables are:

- age: age in years.
- marital: A factor with levels “single”, “married”, and “other”.
- income: yearly income.
- mortgage: A factor with levels “yes” and “no”.
- nr_loans: Number of loans that constomers have.
- risk: A factor with levels “good risk” and “bad risk”.

whitewines White Wines dataset

Description

The whiteWines datasets are related to white variants of the Portuguese “Vinho Verde” wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The dataset can be viewed as classification or regression tasks. the classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Usage

```
df=dsf.load('whitewines')
```

Format

The whiteWines dataset, as a data frame, contains 4898 rows and 12 columns (variables/features).

The 12 variables are:

Input variables (based on physicochemical tests):

- fixed_acidity
- volatile_acidity
- citric_acid
- residual_sugar
- chlorides
- free_sulfur_dioxide
- total_sulfur_dioxide
- density
- pH
- sulphates
- alcohol

Output variable (based on sensory data)

- quality: score between 0 and 10.

Details

This dataset can be downloaded from the UCI machine learning repository:

<https://archive.ics.uci.edu/dataset/186/wine+quality>

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4), 547-553.