Package `dsf`

Description The **dsf** package provides a collection of datasets used in the book `Data Science Foundations and Machine Learning with Python`.

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 Adult dataset

Description

The Adult dataset was collected by the US Census Bureau. The primary task is to predict whether a given adult makes more than $50K per year based on attributes such as education, hours worked per week, and other demographic features. The target variable 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 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".

Source

This dataset is publicly available from UCI Machine Learning Repository: https://archive.ics.uci.edu/dataset/2/adult

Reference

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 organization's social media ad campaign.

Usage

df=dsf.load('advertising')

Format

The advertising dataset 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 whim the add 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.

Source

This dataset is publicly available from Kaggle: https://www.kaggle.com/loveall/clicks-conversion-tracking

bank Bank Marketing dataset

Description This dataset contains data from 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 contains 4521 rows (customers) and 17 columns (variables/features). The 17 variables are explained below.

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".

Source

This dataset is publicly available from UCI Machine Learning Repository: https://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 from Sentient Machine Research 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('caravan')

Format

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

Further information is available at http://www.liacs.nl/~putten/library/cc2000/data.html.

Source

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

This dataset is publicly available from Kaggle: https://www.kaggle.com/datasets/uciml/caravan-insurance-challenge

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 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?).

More information is available at: https://community.amstat.org/stat-computing/data-expo/data-expo-1993

Source

The dataset originates from the 1993 ASA Statistical Graphics Exposition, organized by the American Statistical Association. It is publicly available from DASL: https://dasl.datadescription.com/datafile/cereals/

churn Churn dataset

Description

This synthetic dataset from MLC++ machine learning software is used for modeling customer churn. 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 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/no), whether the customer has a voice mail plan.
- voice_messages: Count, number of voice mail messages.
- intl_plan: Categorical (yes/no), whether the customer has an 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/no), whether the customer has left the company.

Source

This dataset was originally provided by MLC++. The original MLC++ site is no longer available, but the data can be found here:

- OpenML: https://openml.org/d/40701
- data.world: https://data.world/earino/churn
- modeldata package (available on CRAN): https://cran.r-project.org/web/packages/modeldata/index.html

References

Saha, S., Saha, C., Haque, M. M., Alam, M. G. R., & Talukder, A. (2024). ChurnNet: Deep learning enhanced customer churn prediction in telecommunication industry. IEEE access, 12, 4471-4484.

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 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).

Source

This dataset is publicly available from Kaggle: https://www.kaggle.com/sakshigoyal7/credit-card-customers

churntel Telco Customer Churn dataset

Description

Customer *churn* occurs when customers stop doing business with a company, also known as customer attrition. This synthetic dataset from IBM 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 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).

Source

This dataset is publicly available from Kaggle: https://www.kaggle.com/blastchar/telco-customer-churn.

References

https://www.ibm.com/docs/en/cognos-analytics/12.1.0?topic=samples-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 contains 61900 rows and 12 columns (variables/features).

More information is available at: https://data.europa.eu/data/datasets/covid-19-coronavirus-data-daily-up-to-14-december-2020

Source

This dataset is publicly available from the European Union's 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 from ggplot2 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 diamonds dataset contains 53940 rows and 10 columns (variables/features).

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).

Source

This dataset is publicly available in the ggplot2 R package: https://ggplot2.tidyverse.org/reference/diamonds.html

For more information related to the dataset see:

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

drug Drug Classification dataset

Description

A synthetically generated dataset of 200 patients that includes their age, sodium-to-potassium (Na/K) ratio, and the prescribed drug type.

Usage

df=dsf.load('drug')

Format

The drug dataset contains 200 rows and 3 columns (variables/features).

The 3 variables are:

- age: patient age,
- ratio: sodium-to-potassium (Na/K) ratio,
- type: prescribed drug type.

Source

This dataset is generated for the book 'Data Science Foundations and Machine Learning Using Python'.

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 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).

Source

The data is publicly available from Kaggle: https://www.kaggle.com/quantbruce/real-estate-price-prediction

houseprice HousePrice dataset

Description

This dataset, created by Dean De Cock, contains 1460 rows and 81 columns (features). The `saleprice` column is the target.

Usage

df=dsf.load('houseprice')

Format

The housePrice dataset contains 1460 rows and 81 columns (variables/features). More information about the variables can be found at: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Source

The data is publicly available from Kaggle: 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 synthetic insurance dataset 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.

A detailed description of the dataset can be found at: https://www.kaggle.com/mirichoi0218/insurance

Source

The dataset is publicly available from Github: https://github.com/stedy/Machine-Learning-with-R-datasets

Reference

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 contains 40 rows and 8 columns (variables/features).

The 8 variables are:

- spend: daily send of money on PPC (apy-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.

Source

The dataset by Chris Bow is publicly available from Github: https://github.com/chrisBow/marketing-regression-part-one

mpg Auto MPG dataset

Description

The Auto MPG dataset contains information on various car models from the 1970s and 1980s, with the goal of predicting fuel efficiency (miles per gallon, `mpg`). It includes attributes describing engine characteristics, vehicle weight, performance, model year, origin, and car name.

Usage

df=dsf.load('mpg')

Format

The Auto MPG dataset contains 398 observations (cars) and 9 columns (variables/features):

- mpg: miles per gallon, a continuous variable measuring fuel efficiency.
- cylinders: number of cylinders in the engine, a discrete factor with typical values 3, 4, 5, 6, 8.
- displacement: engine displacement in cubic inches, a continuous variable representing engine size.
- horsepower: engine power in horsepower, a continuous variable (may have missing values).
- weight: vehicle weight in pounds, a continuous variable.
- acceleration: time to accelerate from 0 to 60 mph in seconds, a continuous variable.
- model_year: year of the car model, a discrete variable usually coded as two digits (e.g., 70 = 1970).
- origin: origin of the car, a factor with 3 levels (1 = USA, 2 = Europe, 3 = Japan).
- name: car model name, a string variable for identification.

Source

This dataset is publicly available from the UCI machine learning repository: https://archive.ics.uci.edu/dataset/9/auto+mpg

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 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.

Source

The dataset is publicly available from the UCI machine learning repository: https://archive.ics.uci.edu/dataset/186/wine+quality

Reference

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 contains 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 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".

Source

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

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 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.

Source

The dataset is publicly available 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.