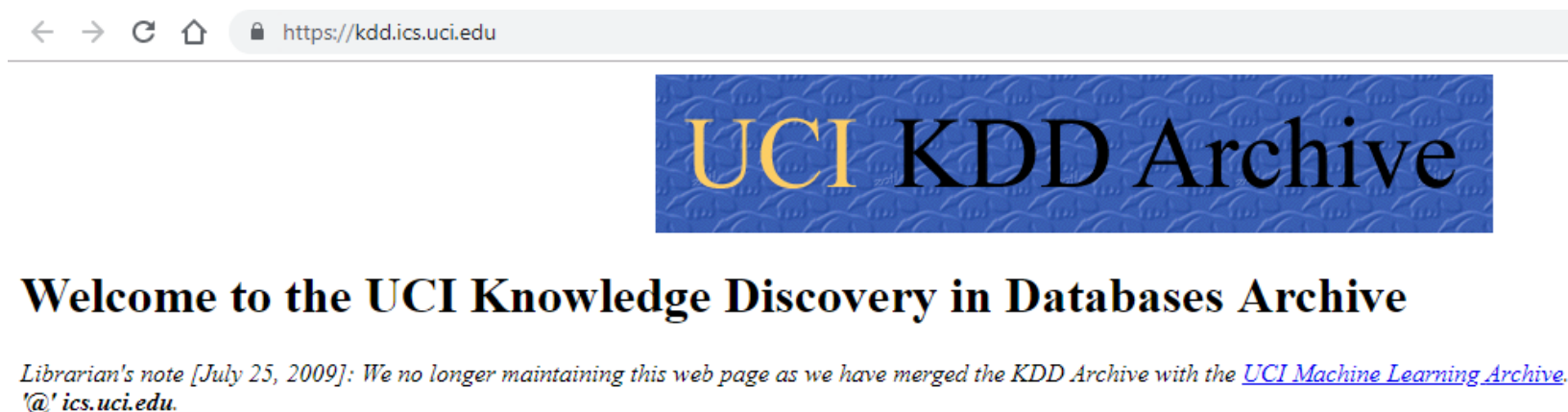


# Fontes de Dados

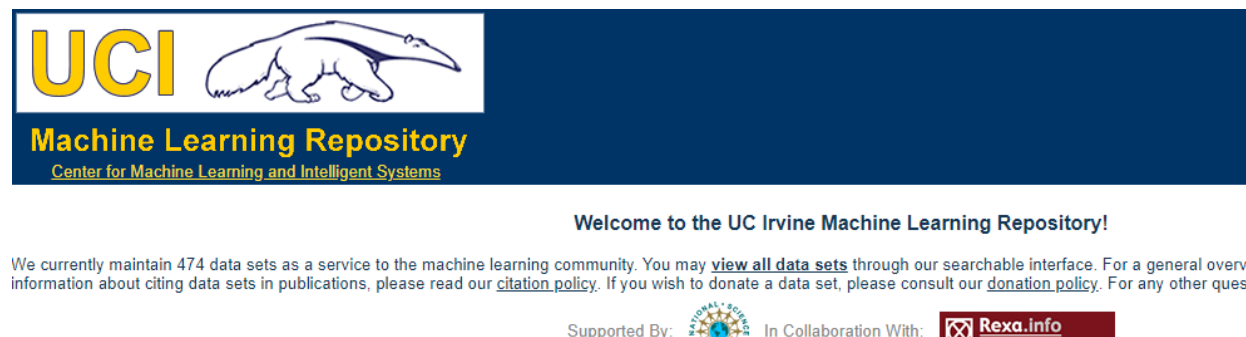


# Fontes de Dados

- Para estudo das técnicas e/ou comparação de algoritmos:
  - **UCI Knowledge Discovery in Databases**  
<https://kdd.ics.uci.edu/>



- **UC Irvine Machine Learning Repository**  
<https://kdd.ics.uci.edu/>



# UCI

## Newest Data Sets:

|             |   |   |
|-------------|---|---|
| 05-07-2019: |    | <a href="#">Metro Interstate Traffic Volume</a>                     |
| 04-22-2019: |    | <a href="#">Facebook Live Sellers in Thailand</a>                   |
| 04-15-2019: |    | <a href="#">Gas sensor array temperature modulation</a>             |
| 04-14-2019: |    | <a href="#">Rice Leaf Diseases</a>                                  |
| 04-10-2019: |    | <a href="#">Parkinson Dataset with replicated acoustic features</a> |
| 04-08-2019: |    | <a href="#">Labeled Text Forum Threads Dataset</a>                  |
| 01-07-2019: |   | <a href="#">EMG data for gestures</a>                               |
| 01-02-2019: |  | <a href="#">Parking Birmingham</a>                                  |
| 12-19-2018: |  | <a href="#">Tarvel Review Ratings</a>                               |
| 12-19-2018: |  | <a href="#">Travel Reviews</a>                                      |

## Most Popular Data Sets (hits since 2007):

|          |   |  |
|----------|---|--|
| 2662543: |    | <a href="#">Iris</a>   |
| 1512870: |    | <a href="#">Adult</a>  |
| 1169782: |    | <a href="#">Wine</a>   |
| 993532:  |    | <a href="#">Car Evaluation</a>                               |
| 958075:  |    | <a href="#">Wine Quality</a>                                 |
| 942275:  |    | <a href="#">Heart Disease</a>                                |
| 941591:  |   | <a href="#">Breast Cancer Wisconsin (Diagnostic)</a>         |
| 907223:  |  | <a href="#">Bank Marketing</a>                               |
| 839495:  |  | <a href="#">Human Activity Recognition Using Smartphones</a> |
| 791423:  |  | <a href="#">Abalone</a>                                      |

# Machine Learning Repository

Center for Machine Learning and Intelligent Systems

UCI

Browse Through: 474 Data Sets

## Default Task

[Classification](#) (353)  
[Regression](#) (98)  
[Clustering](#) (85)  
[Other](#) (55)

## Attribute Type

[Categorical](#) (38)  
[Numerical](#) (311)  
[Mixed](#) (55)

## Data Type

[Multivariate](#) (361)  
[Univariate](#) (23)  
[Sequential](#) (48)  
[Time-Series](#) (93)  
[Text](#) (54)  
[Domain-Theory](#) (23)  
[Other](#) (21)

## Area

[Life Sciences](#) (108)  
[Physical Sciences](#) (49)  
[CS / Engineering](#) (172)  
[Social Sciences](#) (26)  
[Business](#) (30)  
[Game](#) (10)  
[Other](#) (74)

## # Attributes

[Less than 10](#) (115)  
[10 to 100](#) (213)  
[Greater than 100](#) (84)

## Name



[Abalone](#)



[Adult](#)



[Annealing](#)



[Anonymous Microsoft Web Data](#)



[Arrhythmia](#)



[Artificial Characters](#)



[Audiology \(Original\)](#)



# WEKA

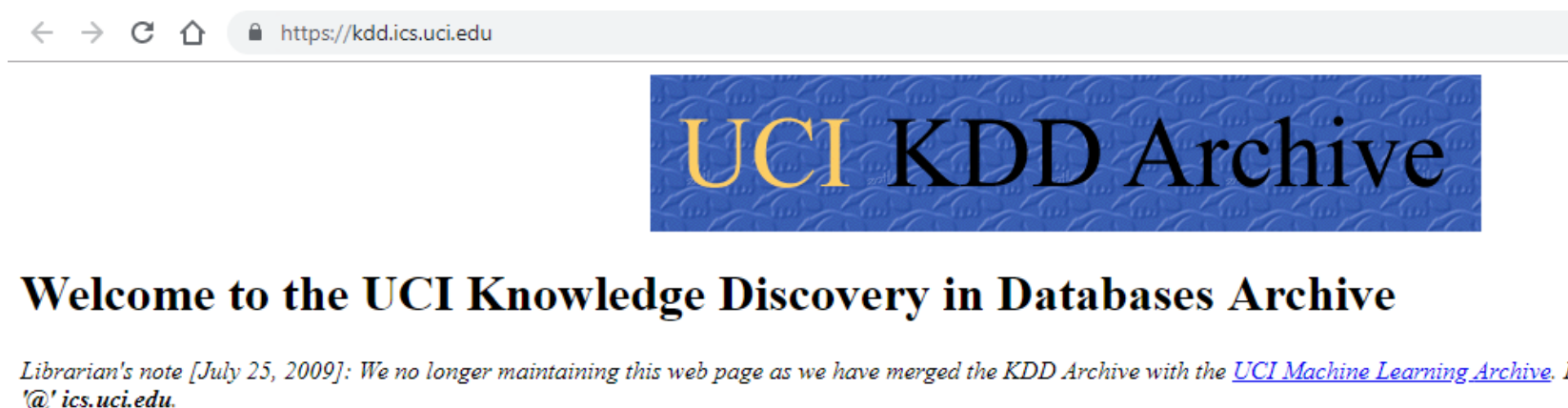
Available separately:

- A jarfile containing 37 classification problems, originally obtained from the **UCI repository** (**datasets-UCI.jar, 1,190,961 Bytes**).
- A jarfile containing 37 regression problems, obtained from various sources (**datasets-numeric.jar, 169,344 Bytes**).
- A jarfile containing 6 agricultural datasets obtained from agricultural researchers in New Zealand (**agridatasets.jar, 31,200 Bytes**).
- A jarfile containing 30 regression datasets collected by Luis Torgo (**regression-datasets.jar, 10,090,266 Bytes**).
- A gzip'ed tar containing **UCI** and **UCI KDD** datasets (**uci-20070111.tar.gz, 17,952,832 Bytes**)
- A gzip'ed tar containing **StatLib** datasets (**statlib-20050214.tar.gz, 12,785,582 Bytes**)
- A gzip'ed tar containing ordinal, real-world datasets donated by **Dr. Arie Ben David** (Holon Inst. of Technology/Israel) (**datasets-arie\_ben\_david.tar.gz, 11,348 Bytes**)
- A zip file containing 19 multi-class (1-of-n) text datasets donated by **George Forman/Hewlett-Packard Labs** (**19MclassTextWc.zip, 14,084,828 Bytes**)
- A bzip'ed tar file containing the Reuters21578 dataset split into separate files according to the ModApte split (**reuters21578-ModApte.tar.bz2, 81,745,032 Bytes**)
- A zip file containing 41 drug design datasets formed using the Adriana.Code software - **www.molecular-networks.com/software/adrianacode** - donated by **Dr. M. Fatih Amasyali** (Yildiz Technical University) (**Drug-datasets.zip, 11,376,153 Bytes**)
- A zip file containing 80 artificial datasets generated from the Friedman function donated by **Dr. M. Fatih Amasyali** (Yildiz Technical University) (**Friedman-datasets.zip, 5,802,204 Bytes**)

Link: <https://www.cs.waikato.ac.nz/ml/weka/datasets.html>


# Fontes de Dados

- Para estudo das técnicas e/ou comparação de algoritmos:



# Bases de dados para o R


www.rdatamining.com/resources/data

**RDataMining.com: R and Data Mining**

[Home](#)  
[Resource News](#)  
[Seminar and Conference News](#)  
[Job News](#)  
▼ [Training](#)  
    [R and Data Mining Course](#)  
    [Tutorial at AusDM 2018](#)  
    [Tutorial at Melbourne Data Science Week](#)  
    [Short Course at University of Canberra](#)  
    [Machine Learning 102](#)  
    [Workshop at SP Jain](#)  
    [Past Trainings and Talks](#)  
▼ [Documents](#)  
    [Introduction to Data Mining with R](#)  
    [R Reference Card for Data Mining](#)  
    [R and Data Mining: Examples and Case Studies](#)  
    [Introduction to Data Mining with R and Data Import/Export in R](#)

[Resources >](#)  
**Free Datasets**

Learn to build your first **vaadin}>** app for free



There are many datasets available online for free for research use. Some of them

If you'd like to have some datasets added to the page, please feel free to send the [yanchang\(at\)RDataMining.com](mailto:yanchang(at)RDataMining.com). Thanks.

- [Geocoded National Address File \(G-NAF\)](#)  
more than 13 million Australian physical address records with geocodes
- [The GeoNames geographical database](#)  
covers all countries and contains over eight million place names, which can geocode for countries, cities, suburbs, places and postcodes.
- [Airport, airline and route data](#)  
6977 airports, 5888 airlines and 59036 routes spanning the globe
- [GDELT: Global Data on Events, Location and Tone](#)  
containing over 200-million geolocated events for 1979 to present






Fonte: <http://www.rdatamining.com/resources/data>

# kaggle

Kaggle, uma subsidiária da Google LLC, é uma comunidade on-line de cientistas de dados e profissionais de aprendizado de máquina.

- Datasets  
<https://www.kaggle.com/datasets>
- Intro to Machine Learning  
<https://www.kaggle.com/learn/intro-to-machine-learning>

PUBLIC

|   |   |
|---|---|
|    | <b>arXiv Dataset</b><br>Cornell University <a href="#">Link</a><br>8 days 877 MB 8.8 1 File (JSON) 3 Tasks            |
|    | <b>All Space Missions from 1957</b><br>Agirlcoding<br>9 days 101 KB 8.5 1 File (CSV) 1 Task                           |
|    | <b>Handwriting Recognition</b><br>landlord<br>17 days 1 GB 9.4 413704 Files (other, CSV)                              |
|   | <b>Sales of summer clothes in E-commerce Wish</b><br>Jeffrey Mvutu Mabilama<br>16 days 351 KB 9.7 1 File (CSV) 1 Task |
|  | <b>Heart Failure Prediction</b><br>Level  |











Kaggle, uma subsidiária da Google LLC, é uma comunidade on-line de cientistas de dados e profissionais de aprendizado de máquina

- Datasets

<https://www.kaggle.com/datasets>

- Intro to Machine Learning

<https://www.kaggle.com/learn/intro>

|   |   |      |
|---|---|------|
|    | <b>Chest X-Ray Images (Pneumonia)</b><br>Paul Mooney<br>2 years 2 GB 7.5 5856 Files (other) 1 Task  | 2482 |
|    | <b>Restaurant Recommendation Challenge</b><br>Andriy Samoshyn<br>a month 534 MB 9.1 10 Files (CSV, other) 1 Task                                  | 75   |
|    | <b>Netflix Movies and TV Shows</b><br>Shivam Bansal<br>7 months 971 KB 10.0 1 File (CSV) 3 Tasks  | 1568 |
|   | <b>COVID-19 Open Research Dataset Challenge (CORD-19)</b><br>Allen Institute For AI Link<br>4 days 4 GB 8.8 167675 Files (JSON, CSV, other) 17... | 8102 |
|  | <b>Hacker News</b><br>Hacker News<br>2 years 7.1 BigQuery   | 324  |
|  | <b>ASHRAE Global Thermal Comfort Database II</b><br>Clayton   | 20   |

# Para fins de enriquecimento...

- Portal Brasileiro de Dados Abertos  
<http://dados.gov.br/>
- IBGE  
<https://www.ibge.gov.br/>
- Datasus  
<http://datasus.saude.gov.br/>  
<http://www2.datasus.gov.br/DATASUS/index.php?area=02>
- Open Datasus  
<https://opendatasus.saude.gov.br/>
- Dados Abertos – Ministério da Agricultura, Pecuária e Abastecimento  
<http://www.agricultura.gov.br/aceso-a-informacao/dadosabertos>  
<http://www.agricultura.gov.br/agroestatisticas/estatisticas-e-dados-basicos-de-economia-agricola>
- Estatísticas da CNI  
<http://www.portaldaindustria.com.br/cni/estatisticas/>
- Dados do Ministério do Trabalho  
[trabalho.gov.br/dados-abertos](http://trabalho.gov.br/dados-abertos)
- Comissão de Valores Mobiliários  
[http://www.cvm.gov.br/menu/aceso\\_informacao/dadosabertos/dadosabertos.html](http://www.cvm.gov.br/menu/aceso_informacao/dadosabertos/dadosabertos.html)
- Turismo  
<http://dados.turismo.gov.br/>
- Dados da Educação  
<http://inep.gov.br/dados>
- Dados Abertos - Produtividade e Comércio Exterior  
[www.mdic.gov.br](http://www.mdic.gov.br)
- Bolsas de Valores  
[http://www.b3.com.br/pt\\_br/market-data-e-indices/servicos-de-dados/market-data/consultas/boletim-diario/arquivos-para-download/](http://www.b3.com.br/pt_br/market-data-e-indices/servicos-de-dados/market-data/consultas/boletim-diario/arquivos-para-download/)

# Observações

- Dados / metadados
- Big Data
- Small Data
- Micro data