Heart Disease Data Set

Download: Data Folder, Data Set Description

Abstract: 4 databases: Cleveland, Hungary, Switzerland, and the VA Long Beach

Data Set Characteristics:	Multivariate	**Number of Instances:**	303	**Area:**	Life
Attribute Characteristics:	Categorical, Integer, Real	**Number of Attributes:**	75	**Date Donated**	1988-07-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	101862

Source:

Creators:

1. Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.

Donor:

David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Data Set Information:

This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by ML researchers to
this date. The "goal" field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Cleveland database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0).

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

One file has been "processed", that one containing the Cleveland database. All four unprocessed files also exist in this directory.

To see Test Costs (donated by Peter Turney), please see the folder "Costs"

To see Test Costs (donated by Peter Turney), please see the folder "Costs"

Attribute Information:

Only 14 attributes used:
1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

Complete attribute documentation:
1 id: patient identification number
2 ccf: social security number (I replaced this with a dummy value of 0)
3 age: age in years
4 sex: sex (1 = male; 0 = female)
5 painloc: chest pain location (1 = substernal; 0 = otherwise)
6 painexer (1 = provoked by exertion; 0 = otherwise)
7 relrest (1 = relieved after rest; 0 = otherwise)
8 pncaden (sum of 5, 6, and 7)
9 cp: chest pain type
-- Value 1: typical angina
-- Value 2: atypical angina
-- Value 3: non-anginal pain
-- Value 4: asymptomatic
10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)
11 htn
12 chol: serum cholestoral in mg/dl
13 smoke: I believe this is 1 = yes; 0 = no (is or is not a smoker)
14 cigs (cigarettes per day)
15 years (number of years as a smoker)
16 fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17 dm (1 = history of diabetes; 0 = no such history)
18 famhist: family history of coronary artery disease (1 = yes; 0 = no)
19 restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
20 ekgmo (month of exercise ECG reading)
21 ekgday(day of exercise ECG reading)
22 ekgyr (year of exercise ECG reading)
23 dig (digitalis used furing exercise ECG: 1 = yes; 0 = no)
24 prop (Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25 nitr (nitrates used during exercise ECG: 1 = yes; 0 = no)
26 pro (calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27 diuretic (diuretic used used during exercise ECG: 1 = yes; 0 = no)
28 proto: exercise protocol
1 = Bruce
2 = Kottus
3 = McHenry
4 = fast Balke
5 = Balke
6 = Noughton
7 = bike 150 kpa min/min (Not sure if "kpa min/min" is what was written!)
8 = bike 125 kpa min/min
9 = bike 100 kpa min/min
10 = bike 75 kpa min/min
11 = bike 50 kpa min/min
12 = arm ergometer

12 = arm ergometer
29 thaldur: duration of exercise test in minutes
30 thaltime: time when ST measure depression was noted

31 met: mets achieved
32 thalach: maximum heart rate achieved
33 thalrest: resting heart rate
34 tpeakbps: peak exercise blood pressure (first of 2 parts)
35 tpeakbpd: peak exercise blood pressure (second of 2 parts)
36 dummy
37 trestbpd: resting blood pressure
38 exang: exercise induced angina (1 = yes; 0 = no)
39 xhypo: (1 = yes; 0 = no)
40 oldpeak = ST depression induced by exercise relative to rest
41 slope: the slope of the peak exercise ST segment
-- Value 1: upsloping
-- Value 2: flat
-- Value 3: downsloping
42 rldv5: height at rest
43 rldv5e: height at peak exercise
44 ca: number of major vessels (0-3) colored by flourosopy
45 restckm: irrelevant
46 exerckm: irrelevant
47 restef: rest raidonuclid (sp?) ejection fraction
48 restwm: rest wall (sp?) motion abnormality
0 = none
1 = mild or moderate
2 = moderate or severe
3 = akinesis or dyskmem (sp?)
49 exeref: exercise radinalid (sp?) ejection fraction
50 exerwm: exercise wall (sp?) motion
51 thal: 3 = normal; 6 = fixed defect; 7 = reversable defect
52 thalsev: not used
53 thalpul: not used
54 earlobe: not used
55 cmo: month of cardiac cath (sp?) (perhaps "call")
56 cday: day of cardiac cath (sp?)
57 cyr: year of cardiac cath (sp?)
58 num: diagnosis of heart disease (angiographic disease status)
-- Value 0: < 50% diameter narrowing
-- Value 1: > 50% diameter narrowing
(in any major vessel: attributes 59 through 68 are vessels)
59 lmt
60 ladprox
61 laddist
62 diag
63 cxmain
64 ramus
65 om1
66 om2
67 rcaprox
68 rcadist
69 lvx1: not used
70 lvx2: not used
71 lvx3: not used
72 lvx4: not used
73 lvf: not used
74 cathef: not used
75 junk: not used
76 name: last name of patient (I replaced this with the dummy string "name")

Relevant Papers:

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.
[Web Link]

David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database." [Web Link]

Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61. [Web Link]

Papers That Cite This Data Set[1]:

Kaizhu Huang and Haiqin Yang and Irwin King and Michael R. Lyu and Laiwan Chan. Biased Minimax Probability Machine for Medical Diagnosis. AMAI. 2004. [View Context].

Zhi-Hua Zhou and Yuan Jiang. NeC4.5: Neural Ensemble Based C4.5. IEEE Trans. Knowl. Data Eng, 16. 2004. [View Context].

Xiaoyong Chai and Li Deng and Qiang Yang and Charles X. Ling. Test-Cost Sensitive Naive Bayes Classification. ICDM. 2004. [View Context].

Jeroen Eggermont and Joost N. Kok and Walter A. Kosters. Genetic Programming for data classification: partitioning the search space. SAC. 2004. [View Context].

Remco R. Bouckaert and Eibe Frank. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. PAKDD. 2004. [View Context].

Gavin Brown. Diversity in Neural Network Ensembles. The University of Birmingham. 2004. [View Context].

David Page and Soumya Ray. Skewing: An Efficient Alternative to Lookahead for Decision Tree Induction. IJCAI. 2003. [View Context].

Jinyan Li and Limsoon Wong. Using Rules to Analyse Bio-medical Data: A Comparison between C4.5 and PCL. WAIM. 2003. [View Context].

Yuan Jiang Zhi and Hua Zhou and Zhaoqian Chen. Rule Learning based on Neural Network Ensemble. Proceedings of the International Joint Conference on Neural Networks. 2002. [View Context].

Baback Moghaddam and Gregory Shakhnarovich. Boosted Dyadic Kernel Discriminants. NIPS. 2002. [View Context].

Peter L. Hammer and Alexander Kogan and Bruno Simeone and Sandor Szedm'ak. R u t c o r Research R e p o r t. Rutgers Center for Operations Research Rutgers University. 2001. [View Context].

Thomas Melluish and Craig Saunders and Ilia Nouretdinov and Volodya Vovk and Carol S. Saunders and I. Nouretdinov V.. The typicalness framework: a comparison with the Bayesian approach. Department of Computer Science. 2001. [View Context].

Robert Burbidge and Matthew Trotter and Bernard F. Buxton and Sean B. Holden. STAR - Sparsity through Automated Rejection. IWANN (1). 2001. [View Context].

Kristin P. Bennett and Ayhan Demiriz and John Shawe-Taylor. A Column Generation Algorithm For Boosting. ICML. 2000. [View Context].

Endre Boros and Peter Hammer and Toshihide Ibaraki and Alexander Kogan and Eddy Mayoraz and Ilya B. Muchnik. An Implementation of Logical Analysis of Data. IEEE Trans. Knowl. Data Eng, 12. 2000. [View Context].

Rudy Setiono and Wee Kheng Leow. FERNN: An Algorithm for Fast Extraction of Rules from Neural Networks. Appl. Intell, 12. 2000. [View Context].

Thomas G. Dietterich. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning, 40. 2000. [View Context].

Lorne Mason and Peter L. Bartlett and Jonathan Baxter. Improved Generalization Through Explicit Optimization of Margins. Machine Learning, 38. 2000. [View Context].

Petri Kontkanen and Petri Myllym and Tomi Silander and Henry Tirri and Peter Gr. On predictive distributions and Bayesian

Petri Kontkanen and Petri Myllym and Tomi Silander and Henry Tirri and Peter Gr. On predictive distributions and Bayesian networks. Department of Computer Science, Stanford University. 2000. [View Context].

Kai Ming Ting and Ian H. Witten. Issues in Stacked Generalization. J. Artif. Intell. Res. (JAIR, 10. 1999. [View Context].

Yoav Freund and Lorne Mason. The Alternating Decision Tree Learning Algorithm . ICML. 1999. [View Context].

Jinyan Li and Xiuzhen Zhang and Guozhu Dong and Kotagiri Ramamohanarao and Qun Sun. Efficient Mining of High Confidience Association Rules without Support Thresholds . PKDD. 1999. [View Context].

Iñaki Inza and Pedro Larrañaga and Basilio Sierra and Ramon Etxeberria and Jose Antonio Lozano and Jos Manuel Peña. Representing the behaviour of supervised classification learning algorithms by Bayesian networks. Pattern Recognition Letters, 20. 1999. [View Context].

Chun-Nan Hsu and Hilmar Schuschel and Ya-Ting Yang. The ANNIGMA-Wrapper Approach to Neural Nets Feature Selection for Knowledge Discovery and Data Mining. Institute of Information Science. 1999. [View Context].

Rudy Setiono and Huan Liu. NeuroLinear: From neural networks to oblique decision rules. Neurocomputing, 17. 1997. [View Context].

Pedro Domingos. Control-Sensitive Feature Selection for Lazy Learners. Artif. Intell. Rev, 11. 1997. [View Context].

Igor Kononenko and Edvard Simec and Marko Robnik-Sikonja. Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF. Appl. Intell, 7. 1997. [View Context].

Floriana Esposito and Donato Malerba and Giovanni Semeraro. A Comparative Analysis of Methods for Pruning Decision Trees. IEEE Trans. Pattern Anal. Mach. Intell, 19. 1997. [View Context].

Jan C. Bioch and D. Meer and Rob Potharst. Bivariate Decision Trees. PKDD. 1997. [View Context].

D. Randall Wilson and Roel Martinez. Machine Learning: Proceedings of the Fourteenth International Conference, Morgan. In Fisher. 1997. [View Context].

. Prototype Selection for Composite Nearest Neighbor Classifiers. Department of Computer Science University of Massachusetts. 1997. [View Context].

Kamal Ali and Michael J. Pazzani. Error Reduction through Learning Multiple Descriptions. Machine Learning, 24. 1996. [View Context].

Peter D. Turney. Cost-Sensitive Classification: Empirical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm . CoRR, csAI/9503102. 1995. [View Context].

Ron Kohavi and Dan Sommerfield. Feature Subset Selection Using the Wrapper Method: Overfitting and Dynamic Search Space Topology. KDD. 1995. [View Context].

Ron Kohavi. The Power of Decision Tables . ECML. 1995. [View Context].

Gabor Melli. A Lazy Model-Based Approach to On-Line Classification . University of British Columbia. 1989. [View Context].

Ayhan Demiriz and Kristin P. Bennett. Chapter 1 OPTIMIZATIONAPPROACHESTOSEMI-SUPERVISED LEARNING. Department of Decision Sciences and Engineering Systems & Department of Mathematical Sciences, Rensselaer Polytechnic Institute. [View Context].

Zhi-Hua Zhou and Xu-Ying Liu. Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem. [View Context].

Ayhan Demiriz and Kristin P. Bennett and John Shawe and I. Nouretdinov V.. Linear Programming Boosting via Column Generation. Dept. of Decision Sciences and Eng. Systems, Rensselaer Polytechnic Institute. [View Context].

Adil M. Bagirov and Alex Rubinov and A. N. Soukhojak and John Yearwood. Unsupervised and supervised data classification via nonsmooth and global optimization. School of Information Technology and Mathematical Sciences, The University of Ballarat. [View Context].

Alexander K. Seewald. Dissertation Towards Understanding Stacking Studies of a General Ensemble Learning Scheme ausgefuhrt zum Zwecke der Erlangung des akademischen Grades eines Doktors der technischen Naturwissenschaften. [View Context].

Wl odzisl/aw Duch and Karol Grudzinski. Search and global minimization in similarity-based methods. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Computer Methods, Nicholas Copernicus University. [View Context].

Wl/odzisl/aw Duch and Karol Grudzinski and Geerd H. F Diercksen. Minimal distance neural methods. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Elena Smirnova and Ida G. Sprinkhuizen-Kuyper and I. Nalbantis and b. ERIM and Universiteit Rotterdam. Unanimous Voting using Support Vector Machines . IKAT, Universiteit Maastricht. [View Context].

Glenn Fung and Sathyakama Sandilya and R. Bharat Rao. Rule extraction from Linear Support Vector Machines . Computer-Aided Diagnosis & Therapy, Siemens Medical Solutions, Inc. [View Context].

John G. Cleary and Leonard E. Trigg. Experiences with OB1, An Optimal Bayes Decision Tree Learner. Department of Computer Science University of Waikato. [View Context].

Chiranjib Bhattacharyya and Pannagadatta K. S and Alexander J. Smola. A Second order Cone Programming Formulation for Classifying Missing Data. Department of Computer Science and Automation Indian Institute of Science. [View Context].

Krista Lagus and Esa Alhoniemi and Jeremias Seppa and Antti Honkela and Arno Wagner. INDEPENDENT VARIABLE GROUP ANALYSIS IN LEARNING COMPACT REPRESENTATIONS FOR DATA . Neural Networks Research Centre, Helsinki University of Technology. [View Context].

Federico Divina and Elena Marchiori. Handling Continuous Attributes in an Evolutionary Inductive Learner. Department of Computer Science Vrije Universiteit. [View Context].

Kristin P. Bennett and Erin J. Bredensteiner. Geometry in Learning. Department of Mathematical Sciences Rensselaer Polytechnic Institute. [View Context].

Bruce H. Edmonds. Using Localised `Gossip' to Structure Distributed Learning. Centre for Policy Modelling. [View Context].

H. -T Lin and C. -J Lin. A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods . Department of Computer Science and Information Engineering National Taiwan University. [View Context].

Wl odzisl and Rafal Adamczak and Krzysztof Grabczewski and Grzegorz Zal. A hybrid method for extraction of logical rules from data. Department of Computer Methods, Nicholas Copernicus University. [View Context].

Liping Wei and Russ B. Altman. An Automated System for Generating Comparative Disease Profiles and Making Diagnoses. Section on Medical Informatics Stanford University School of Medicine, MSOB X215. [View Context].

Adil M. Bagirov and John Yearwood. A new nonsmooth optimization algorithm for clustering. Centre for Informatics and Applied Optimization, School of Information Technology and Mathematical Sciences, University of Ballarat. [View Context].

Ron Kohavi and George H. John. Automatic Parameter Selection by Minimizing Estimated Error. Computer Science Dept. Stanford University. [View Context].

Rafael S. Parpinelli and Heitor S. Lopes and Alex Alves Freitas. PART FOUR: ANT COLONY OPTIMIZATION AND IMMUNE SYSTEMS Chapter X An Ant Colony Algorithm for Classification Rule Discovery . CEFET-PR, Curitiba. [View Context].

Rudy Setiono and Wee Kheng Leow. Generating rules from trained network using fast pruning. School of Computing National University of Singapore. [View Context].

Citation Request:

The authors of the databases have requested that any publications resulting from the use of the data include the names of the principal investigator responsible for the data collection at each institution. They would be:

[1] Papers were automatically harvested and associated with this data set, in collaboration with Rexa.info

2. University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
3. University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.
4. V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert [illegible]

Supported By: National Science Foundation In Collaboration With: Rexa.info