Center for Machine Learning and Intelligent Systems

Pima Indians Diabetes Data Set

Download: Data Folder, Data Set Description

Abstract: From National Institute of Diabetes and Digestive and Kidney Diseases; Includes cost data (donated by Peter Turney)

Data Set Characteristics:	Multivariate	**Number of Instances:**	768	**Area:**	Life
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	8	**Date Donated**	1990-05-09
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	381460

Source:

Original Owners:

National Institute of Diabetes and Digestive and Kidney Diseases

Donor of database:

Vincent Sigillito (vgs '@' aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

Data Set Information:

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

Attribute Information:

1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

9. Class variable (0 or 1)

** **UPDATE:** Until 02/28/2011 this web page indicated that there were no missing values in the dataset. As pointed out by a repository user, this cannot be true: there are zeros in places where they are biologically impossible, such as the blood pressure attribute. It seems very likely that zero values encode missing data. However, since the dataset donors made no such statement we encourage you to use your best judgement and state your assumptions.

Relevant Papers:

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261--265). IEEE Computer Society Press.
[Web Link]

Papers That Cite This Data Set[1]:

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

Eibe Frank and Mark Hall. Visualizing Class Probability Estimators. PKDD. 2003. [View Context].

Michael L. Raymer and Travis E. Doom and Leslie A. Kuhn and William F. Punch. Knowledge discovery in medical and biological datasets using a hybrid Bayes classifier/evolutionary algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 33. 2003. [View Context].

Marina Skurichina and Ludmila Kuncheva and Robert P W Duin. Bagging and Boosting for the Nearest Mean Classifier: Effects of Sample Size on Diversity and Accuracy. Multiple Classifier Systems. 2002. [View Context].

Ilya Blayvas and Ron Kimmel. Multiresolution Approximation for Classification. CS Dept. Technion. 2002. [View Context].

Tao Jiang and Art B. Owen. Quasi-regression for visualization and interpretation of black box functions. Department of Statistics Stanford University. 2002. [View Context].

Peter Sykacek and Stephen J. Roberts. Adaptive Classification by Variational Kalman Filtering. NIPS. 2002. [View Context].

Jochen Garcke and Michael Griebel and Michael Thess. Data Mining with Sparse Grids. Computing, 67. 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].

Simon Tong and Daphne Koller. Restricted Bayes Optimal Classifiers. AAAI/IAAI. 2000. [View Context].

Stavros J. Perantonis and Vassilis Virvilis. Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis. Neural Processing Letters, 10. 1999. [View Context].

Huan Liu and Rudy Setiono. Feature Transformation and Multivariate Decision Tree Induction. Discovery Science. 1998. [View Context].

Thomas G. Dietterich. Approximate Statistical Test For Comparing Supervised Classification Learning Algorithms. Neural Computation, 10. 1998. [View Context].

Kristin P. Bennett and Erin J. Bredensteiner. A Parametric Optimization Method for Machine Learning. INFORMS Journal on Computing, 9. 1997. [View Context].

Jennifer A. Blue and Kristin P. Bennett. Hybrid Extreme Point Tabu Search. Department of Mathematical Sciences Rensselaer Polytechnic Institute. 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].

Wl odzisl/aw Duch and Rudy Setiono and Jacek M. Zurada. Computational intelligence methods for rule-based data understanding. [View Context].

Michalis K. Titsias and Aristidis Likas. Shared Kernel Models for Class Conditional Density Estimation. [View Context].

Lawrence O. Hall and Nitesh V. Chawla and Kevin W. Bowyer. Combining Decision Trees Learned in Parallel. Department of Computer Science and Engineering, ENB 118 University of South Florida. [View Context].

Charles Campbell and Nello Cristianini. Simple Learning Algorithms for Training Support Vector Machines. Dept. of Engineering Mathematics. [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].

Chotirat Ann and Dimitrios Gunopulos. Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Computer Science Department University of California. [View Context].

Federico Divina and Elena Marchiori. Knowledge-Based Evolutionary Search for Inductive Concept Learning. Vrije Universiteit of Amsterdam. [View Context].

Michael Lindenbaum and Shaul Markovitch and Dmitry Rusakov. Selective Sampling Using Random Field Modelling. [View Context].

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

Ilya Blayvas and Ron Kimmel. INVITED PAPER Special Issue on Multiresolution Analysis Machine Learning via Multiresolution Approximation. [View Context].

Andrew Watkins and Jon Timmis and Lois C. Boggess. Artificial Immune Recognition System (AIRS): An ImmuneInspired Supervised Learning Algorithm. (abw5,jt6@kent.ac.uk) Computing Laboratory, University of Kent. [View Context].

Ilya Blayvas and Ron Kimmel. Efficient Classification via Multiresolution Training Set Approximation. CS Dept. Technion. [View Context].

Matthias Scherf and W. Brauer. Feature Selection by Means of a Feature Weighting Approach. GSF - National Research Center for Environment and Health. [View Context].

Rudy Setiono and Huan Liu. Neural-Network Feature Selector. Department of Information Systems and Computer Science National University of Singapore. [View Context].

Christopher P. Diehl and Gert Cauwenberghs. SVM Incremental Learning, Adaptation and Optimization. Applied Physics Laboratory Johns Hopkins University. [View Context].

Citation Request:

Please refer to the Machine Learning Repository's citation policy

---

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