



[Center for Machine Learning and Intelligent Systems](#)

[About](#) [Citation Policy](#) [Donate a Data Set](#)  
[Contact](#)

 

☒ Repository ☐ Web



[View ALL Data Sets](#)

## Adult Data Set

*Download:* [Data Folder](#), [Data Set Description](#)

**Abstract:** Predict whether income exceeds \$50K/yr based on census data.  
 Also known as "Census Income" dataset.



<b>Data Set Characteristics:</b>	Multivariate	<b>Number of Instances:</b>	48842	<b>Area:</b>	Social
<b>Attribute Characteristics:</b>	Categorical, Integer	<b>Number of Attributes:</b>	14	<b>Date Donated</b>	1996-05-01
<b>Associated Tasks:</b>	Classification	<b>Missing Values?</b>	Yes	<b>Number of Web Hits:</b>	414241

### Source:

Donor:

Ronny Kohavi and Barry Becker  
 Data Mining and Visualization  
 Silicon Graphics.  
 e-mail: ronnyk '@' live.com for questions.

### Data Set Information:

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Prediction task is to determine whether a person makes over 50K a year.

### Attribute Information:

Listing of attributes:

>50K, <=50K.

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: continuous.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinidad&Tobago, Peru, Hong, Holand-Netherlands.

## Relevant Papers:

Ron Kohavi, "Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid", Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996

[\[Web Link\]](#)

## Papers That Cite This Data Set<sup>1</sup>:



Rakesh Agrawal and Ramakrishnan Iyengar and Dilys Thomas. [Privacy Preserving OLAP](#). SIGMOD Conference. 2005. [\[View Context\]](#).

Rich Caruana and Alexandru Niculescu-Mizil. [An Empirical Evaluation of Supervised Learning for ROC Area](#). ROCAI. 2004. [\[View Context\]](#).

Rich Caruana and Alexandru Niculescu-Mizil and Geoff Crew and Alex Ksikes. [Ensemble selection from libraries of models](#). ICML. 2004. [\[View Context\]](#).

Bianca Zadrozny. [Learning and evaluating classifiers under sample selection bias](#). ICML. 2004. [\[View Context\]](#).

Wei-Chun Kao and Kai-Min Chung and Lucas Assun and Chih-Jen Lin. [Decomposition Methods for Linear Support Vector Machines](#). Neural Computation, 16. 2004. [\[View Context\]](#).

Saharon Rosset. [Model selection via the AUC](#). ICML. 2004. [\[View Context\]](#).

I. Yonca. [Maximum a Posteriori Tree Augmented Naive Bayes Classifiers](#). O EN INTEL.LIG ` ENCIA ARTIFICIAL CSIC. 2003. [\[View Context\]](#).

Christopher R. Palmer and Christos Faloutsos. Electricity Based External Similarity of Categorical Attributes. PAKDD. 2003. [\[View Context\]](#).

S. Sathiya Keerthi and Chih-Jen Lin. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel. Neural Computation, 15. 2003. [\[View Context\]](#).

Thomas Serafini and G. Zanghirati and Del Zanna and T. Serafini and Gaetano Zanghirati and Luca Zanni. DIPARTIMENTO DI MATEMATICA. Gradient Projection Methods for. 2003. [\[View Context\]](#).

Bart Hamers and J. A. K Suykens. Coupled Transductive Ensemble Learning of Kernel Models. Bart De Moor. 2003. [\[View Context\]](#).

Andrew W. Moore and Weng-Keen Wong. Optimal Reinsertion: A New Search Operator for Accelerated and More Accurate Bayesian Network Structure Learning. ICML. 2003. [\[View Context\]](#).

Alexander J. Smola and Vishy Vishwanathan and Eleazar Eskin. Laplace Propagation. NIPS. 2003. [\[View Context\]](#).

Nitesh V. Chawla and Kevin W. Bowyer and Lawrence O. Hall and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. (JAIR, 16. 2002. [\[View Context\]](#).

S. Sathiya Keerthi and Kaibo Duan and Shirish Krishnaji Shevade and Aun Neow Poo. A Fast Dual Algorithm for Kernel Logistic Regression. ICML. 2002. [\[View Context\]](#).

Ramesh Natarajan and Edwin P D Pednault. Segmented Regression Estimators for Massive Data Sets. SDM. 2002. [\[View Context\]](#).

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. KDD. 2002. [\[View Context\]](#).

Zhiyuan Chen and Johannes Gehrke and Flip Korn. Query Optimization In Compressed Database Systems. SIGMOD Conference. 2001. [\[View Context\]](#).

Stephen D. Bay. Multivariate Discretization for Set Mining. Knowl. Inf. Syst, 3. 2001. [\[View Context\]](#).

Bernhard Pfahringer and Geoffrey Holmes and Richard Kirkby. Optimizing the Induction of Alternating Decision Trees. PAKDD. 2001. [\[View Context\]](#).

Stephen D. Bay and Michael J. Pazzani. Detecting Group Differences: Mining Contrast Sets. Data Min. Knowl. Discov, 5. 2001. [\[View Context\]](#).

Jie Cheng and Russell Greiner. Learning Bayesian Belief Network Classifiers: Algorithms and System. Canadian Conference on AI. 2001. [\[View Context\]](#).

Dmitry Pavlov and Jianchang Mao and Byron Dom. Scaling-Up Support Vector Machines Using Boosting Algorithm. ICPR. 2000. [\[View Context\]](#).

Gary M. Weiss and Haym Hirsh. A Quantitative Study of Small Disjuncts: Experiments and Results. Department of Computer Science Rutgers University. 2000. [\[View Context\]](#).

Dmitry Pavlov and Darya Chudova and Padhraic Smyth. Towards scalable support vector machines using squashing. KDD. 2000. [\[View Context\]](#).

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

Petri Kontkanen and Jussi Lahtinen and Petri Myllymaki and Tomi Silander and Henry Tirri. Proceedings of Pre- and

[Post-processing in Machine Learning and Data Mining: Theoretical Aspects and Applications, a workshop within Machine Learning and Applications](#). Complex Systems Computation Group (CoSCo). 1999. [\[View Context\]](#).

Jie Cheng and Russell Greiner. [Comparing Bayesian Network Classifiers](#). UAI. 1999. [\[View Context\]](#).

Yk Huhtala and Juha Kärkkäinen and Pasi Porkka and Hannu Toivonen. [Efficient Discovery of Functional and Approximate Dependencies Using Partitions](#). ICDE. 1998. [\[View Context\]](#).

John C. Platt. [Using Analytic QP and Sparseness to Speed Training of Support Vector Machines](#). NIPS. 1998. [\[View Context\]](#).

Ron Kohavi. [Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid](#). KDD. 1996. [\[View Context\]](#).

Haixun Wang and Philip S. Yu. [SSDT-NN: A Subspace-Splitting Decision Tree Classifier with Application to Target Selection](#). IBM T. J. Watson Research Center. [\[View Context\]](#).

S. V. N Vishwanathan and Alexander J. Smola and M. Narasimha Murty. [considerably faster than competing methods such as Sequential Minimal Optimization or the Nearest Point Algorithm](#). Machine Learning Program, National ICT for Australia. [\[View Context\]](#).

Grigorios Tsoumakas and Ioannis P. Vlahavas. [Fuzzy Meta-Learning: Preliminary Results](#). Greek Secretariat for Research and Technology. [\[View Context\]](#).

Josep Roure Alcobé. [Incremental Hill-Climbing Search Applied to Bayesian Network Structure Learning](#). Escola Universitria Politcnica de Mataró. [\[View Context\]](#).

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

Chris Giannella and Bassem Sayrafi. [An Information Theoretic Histogram for Single Dimensional Selectivity Estimation](#). Department of Computer Science, Indiana University Bloomington. [\[View Context\]](#).

Rong-En Fan and P. -H Chen and C. -J Lin. [Working Set Selection Using the Second Order Information for Training SVM](#). Department of Computer Science and Information Engineering National Taiwan University. [\[View Context\]](#).

Petri Kontkanen and Jussi Lahtinen and Petri Myllymaki and Tomi Silander and Henry Tirri. [USING BAYESIAN NETWORKS FOR VISUALIZING HIGH-DIMENSIONAL DATA](#). Complex Systems Computation Group (CoSCo). [\[View Context\]](#).

Ahmed Hussain Khan and Intensive Care. [Multiplier-Free Feedforward Networks](#). 174. [\[View Context\]](#).

Luc Hoegaerts and J. A. K Suykens and J. Vandewalle and Bart De Moor. [Subset Based Least Squares Subspace Regression in RKHS](#). Katholieke Universiteit Leuven Department of Electrical Engineering, ESAT-SCD-SISTA. [\[View Context\]](#).

David R. Musicant and Alexander Feinberg. [Active Set Support Vector Regression](#). [\[View Context\]](#).

Luc Hoegaerts and J. A. K Suykens and J. Vandewalle and Bart De Moor. [Primal Space Sparse Kernel Partial Least Squares Regression for Large Scale Problems Special Session paper](#). Katholieke Universiteit Leuven Department of Electrical Engineering, ESAT-SCD-SISTA. [\[View Context\]](#).

Kuan-ming Lin and Chih-Jen Lin. [A Study on Reduced Support Vector Machines](#). Department of Computer Science and Information Engineering National Taiwan University. [\[View Context\]](#).

Luca Zanni. [An Improved Gradient Projection-based Decomposition Technique for Support Vector Machines](#).

Dipartimento di Matematica, Universitdi Modena e Reggio Emilia. [\[View Context\]](#).

Jeff G. Schneider and Andrew W. Moore. [Active Learning in Discrete Input Spaces](#). School of Computer Science Carnegie Mellon University. [\[View Context\]](#).

Omid Madani and David M. Pennock and Gary William Flake. [Co-Validation: Using Model Disagreement to Validate Classification Algorithms](#). Yahoo! Research Labs. [\[View Context\]](#).

Ron Kohavi and Barry G. Becker and Dan Sommerfield. [Improving Simple Bayes](#). Data Mining and Visualization Group Silicon Graphics, Inc. [\[View Context\]](#).

Shi Zhong and Weiyu Tang and Taghi M. Khoshgoftaar. [Boosted Noise Filters for Identifying Mislabeled Data](#). Department of Computer Science and Engineering Florida Atlantic University. [\[View Context\]](#).

David R. Musicant. [DATA MINING VIA MATHEMATICAL PROGRAMMING AND MACHINE LEARNING](#). Doctor of Philosophy (Computer Sciences) UNIVERSITY. [\[View Context\]](#).

William W. Cohen and Yoram Singer. [A Simple, Fast, and Effective Rule Learner](#). AT&T Labs--Research Shannon Laboratory. [\[View Context\]](#).

## 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](#)

Supported By:



In Collaboration With:



[About](#) || [Citation Policy](#) || [Donation Policy](#) || [Contact](#) || [CML](#)