

Adult Data Set

Dataset: <https://archive.ics.uci.edu/ml/datasets/adult>
Data Card Authors: Aryan Chaurasia

Adult income dataset based on 1994 census mainly used to predict whether income exceeds \$50K/year based on census data. Has been cited in more than 40 papers and has more than 2 million web hits.

US Census and Ronny Kohavi and Barry Becker

Publishers: UCI Machine Learning Repository

PUBLISHING ORGANIZATION	INDUSTRY SECTOR	PUBLISHER CONTACT
<i>Write the names of the institution or organization responsible for publishing the dataset.</i>	<i>Bold to select all applicable. 👉 Do not delete any unselected choices.</i>	<i>Provide publisher contact details. For dataset owners, see next row.</i>
Organization Name: University of California, Irvine, School of Information and Computer Sciences	Corporate Academic Not-for-profit Individual Others (please Specify)	<ul style="list-style-type: none">University of California, Irvine, School of Information and Computer SciencesSilicon Graphicsml-repository@ics.uci.edu


Dataset Owners

DATASET TEAM(S)	DATASET CONTACT	DATASET AUTHORS
<i>Write the names of the groups or team(s) that own the dataset.</i>	<i>How can dataset owners be contacted for questions about the model? See previous row for publishing institution.</i>	<i>Write the names of all authors associated with the dataset. Provide the affiliation and year if different from publishing institutions or multiple affiliations:</i>
Ronny Kohavi and Barry Becker	<ul style="list-style-type: none">ronnyk@live.com	<ul style="list-style-type: none">Ronny Kohavi and Barry BeckerData Mining and Visualization,Silicon Graphics, Inc, 1994

Funding Sources

FUNDING INSTITUTION(S)	FUNDING DETAILS
<i>Write the names of the funding institutions.</i>	<i>Provide a short summary of funding sources and other support, including details such as programs or projects that may have funded the creation, collection, or curation of the dataset. Include links to relevant documents where applicable.</i>
Silicon Graphics Inc.	The data was first used in a research article published by the authors while they were working for Silicon Graphics Inc.

Dataset Overview

DATASET SUBJECT	DATASET SNAPSHOT	DESCRIPTION OF CONTENT
<i>Bold to select all applicable.</i>  Do not delete any unselected choices.	<i>Fill out details as indicated, adding rows as needed. Include links to additional table(s) with more detailed breakdowns in the caption.</i>	<i>Provide a short summary of the dataset content. Include links where applicable.</i>
Sensitive Data about people Non-Sensitive Data about people Data about natural phenomena Data about places and objects Synthetically generated data Data about systems or products and their behaviors Unknown Others* (*please specify)	Size of dataset 123456 MB Number of Instances 48842 Number of Fields 15 Labeled Classes 1 Number of Labels 2 Average labels per instance 1	Dataset includes columns like age, work class (private, self-employed and etc.), fnlwgt (Final weight), education (Bachelors, HS), Education-Num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, country and target column with two labels whether their income is less or greater than \$50K.

DESCRIPTIVE STATISTICS

Add basic statistics for each field here, as relevant. If there is insufficient space, focus on the most important or critical fields for this dataset. E.g., some statistics will be relevant for numeric data, but not for strings.

Statistic	Age	fnlwgt	Education-Num	Capital Gain	Capital Loss	Hours per week
count	45222	4.522200e+04	45222	45222	45222	45222
mean	38.547941	1.897347e+05	10.118460	1101.430344	88.595418	40.938017
std	13.217870	1.056392e+05	2.552881	7506.430084	404.956092	12.007508
min	17.00	1.349200e+04	1.00	0	0	1.00
25%	28.00	1.173882e+05	9.00	0	0	40.00
50%	37.00	1.783160e+05	10.00	0	0	40.00
75%	47.00	2.379260e+05	13.00	0	0	45.00
max	90.00	1.490400e+06	16.00	99999.00	4356.00	99.00
mode	36	203488	9	0	0	40

Caption for table: Provide links to extended tables where relevant.

Sensitivity of Data		
SENSITIVE DATA	FIELDS WITH SENSITIVE DATA	SECURITY AND PRIVACY HANDLING
<i>Bold to select all applicable.</i>  Do not delete any unselected choices.	<i>Please indicate which features or fields might contain sensitive or personally identifiable information, and if or not collection was intentional using the format below:</i>	<i>Provide a short summary of measures or steps to handle sensitive data in this dataset. Include links and metrics where applicable.</i>
User Content User Metadata User Activity Data Identifiable Data Sensitive Data Business Data Employee Data Pseudonymous Data Anonymous Data Health Data Children’s Data None Others* (*please specify)	<div>Intentionally Collected Sensitive Data Age: Anonymous Data Education: Anonymous Data Salary: Anonymous Data Occupation: Anonymous Data Sex: Anonymous Data Relationship: Anonymous Data Marital Status: Anonymous Data Work class: Anonymous Data</div> <div>Unintentionally Collected Sensitive Data None</div>	The dataset contains no personally identifiable information. The dataset has all anonymous data.
	RELEVANT LINKS	RISKS AND MITIGATIONS
	<i>Provide link(s) to documents that describe any S/PII where available:</i>	<i>Provide a short summary of how risks from PII or sensitive information have been mitigated in the dataset. Include links and metrics where applicable.</i>
	<ul style="list-style-type: none">N/A as there is no personally identifiable information.	<ul style="list-style-type: none">N/A as there is no personally identifiable information.
Dataset Version and Maintenance		
VERSION STATUS	DATASET VERSION	MAINTENANCE PLAN
<i>Bold to select ONE.</i>  Do not delete any unselected choices.	<i>Provide details about this version of the dataset.</i>	<i>Provide a short summary of how the dataset is maintained, including information about refreshes, versioning criteria, errors, feedback and/or recourse. Include links and metrics where applicable.</i>

<p>Regularly Updated</p> <p>New versions of the dataset have been or will continue to be made available.</p> <p>Actively Maintained</p> <p>No new versions will be made available, but this dataset will be actively maintained, including but not limited to updates to the data.</p> <p>Limited Maintenance</p> <p>The data will not be updated, but any technical issues will be addressed.</p> <p>Deprecated</p> <p>This dataset is obsolete or is no longer being maintained.</p>	<p>Current Version 1.0</p> <p>Last Updated 05/1996</p> <p>Release Date 05/1996</p>	<p>Deprecated and hence no longer maintained. However, is publicly available and is popular for classification models.</p>
	NEXT PLANNED UPDATE	EXPECTED UPDATES OR CRITERIA
<p>⚠ Fill this if this dataset is</p> <p>(a) Regularly updated</p> <p>(b) Actively maintained and another version is planned</p>	<p>Provide details about the next planned update.</p>	<p>Provide a short summary for readers to understand updates to the dataset and/or data. Include links, charts, and visualizations as appropriate.</p>
	<ul style="list-style-type: none">Not Available	<ul style="list-style-type: none">Not Available


Example of Data Points

PRIMARY DATA MODALITY	SAMPLING OF DATA POINTS	DATA FIELDS
<i>Bold to select ONE (primary modality).</i> <i>👉 Do not delete any unselected choices.</i>	<i>Link to multiple data points or exploratory demos. If access is restricted, consider adding a fake example that provides a realistic description of data points in the dataset.</i>	<i>Provide a list of fields in data points, including a description and notes on how to interpret fields in an example of data in this dataset.</i>
Image Data Text Data Tabular Data Audio Data Video Data Time Series Graph Data Geospatial Data Multimodal (Please specify) Others (please specify) Unknown	<ul style="list-style-type: none">• Link to demo: https://epistasislab.github.io/pmlb/profile/adult.html• Link to a typical or within normal distribution example: https://epistasislab.github.io/pmlb/profile/adult.html• Link to an outlier or out-of-distribution example: https://epistasislab.github.io/pmlb/profile/adult.html	<ul style="list-style-type: none">• A field will include an anonymous person's age along with other data like their work class (private, self-employed and etc.), fnlwgt (Final weight), education (Bachelors, HS), Education-Num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, country and target column with two labels whether their income is less or greater than \$50K.
	EXAMPLE: TYPICAL DATA POINT	EXAMPLE: OUTLIER DATA POINT
	<i>Copy-and-paste a typical data point. Include a description of what makes it typical. If restricted, consider adding a fake example that provides a realistic description of data points in the dataset.</i>	<i>Copy-and-paste an atypical or outlier data point. Include a description of what makes it atypical. If restricted, consider adding a fake example that provides a realistic description of data points in the dataset.</i>


	<p><write description and what makes this datapoint typical here></p> <p>E.g. of Data Point:</p> <div>en,1,1-1," An anonymous male who is 50 years old, who has 13 years of education and his occupations is stated as Handlers-cleaners, he identifies his race as black, and his native country as the United States of America. He has never married works 13 hours a week and has declared his capital loss and gains as 0 and has also stated that he earns less than 50K a year. "</div>	<p><write description and what makes this datapoint atypical here></p> <p>E.g. of Data Point:</p> <div>en,1,1- 48840," An anonymous male who is 44 years old, who has 13 years of education and his occupation is stated as Craft-repair, he identifies his race as black, and his native country as the United States of America. He is married and works 40 hours a week and has declared his capital loss as 0 and gains as 5455 has also stated that he earns more than 50K a year."</div>
--	--	--

Motivations & Use

Motivations

DATASET PURPOSE(S)	KEY DOMAINS AND APPLICATION(S)	PRIMARY MOTIVATION(S)
<i>Bold to select ONE.</i>  Do not delete any unselected choices.	<i>Use comma-separated tags to indicate the key domains for this dataset.</i>	<i>List the primary motivations for creating or curating this dataset:</i>
Monitoring Research Production Others (please specify)	Domains Machine Learning, Classification Problem Space Income prediction accuracy based on gender.	E.g. <ul style="list-style-type: none">• Create machine learning models to predict income• Encourage academics to come up with better techniques for prediction

Intended Use

DATASET USAGE	INTENDED AND/OR SUITABLE USE CASE(S)	UNSUITABLE USE CASE(S)
<i>Bold to select ONE.</i>  Do not delete any unselected choices.	<i>Summarize the intended and known use cases of this dataset:</i>	<i>Summarize any known problematic use cases of this dataset:</i>
Safe for production use Safe for research use Conditional use- some unsafe applications Only approved use Others (please specify)	<ul style="list-style-type: none">• Research Work for testing new techniques• Education mostly to build ML model for the dataset.	<ul style="list-style-type: none">• Using model trained on this data on real world application where decisions made by model would have real life consequences.
	PROBLEM SPACE AND RESEARCH QUESTIONS(S)	PUBLICATION GUIDELINES
	<i>Describe the specific problem space that this dataset intends to address. Include any specific research questions.</i>	<i>Include any guidelines and steps for citing this dataset in research and/or production work.</i>
	<ul style="list-style-type: none">• Scaling up accuracy of classifier.	<ul style="list-style-type: none">• Since the dataset is donated to UCI machine learning repository one must acknowledge UCI machine learning repository and also the authors of the dataset.

Access, Retention, & Wipeout

Access

ACCESS TYPE	DOCUMENTATION LINKS	ACCESS PREREQUISITES
Bold to select ONE. 👉 Do not delete any unselected choices.	Provide links that describe documentation to access this dataset:	Please describe any required training or prerequisites to access: this dataset.
Unrestricted Conditional Open Access Others (please specify)	<ul style="list-style-type: none">Website: https://archive.ics.uci.edu/ml/datasets/adult	<ul style="list-style-type: none">Not applicable
	DIRECT LINKS TO DATASET	ACCESS POLICY
	Provide links to access this dataset:	Summarize the access policy associated with this dataset. Use this space to include any other information or links that might be relevant to accessing the dataset.
	<ul style="list-style-type: none">Direct download link: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.dataOther repository: https://www.kaggle.com/datasets/wenruihu/adult-income-dataset <p>Code to download data</p> <pre># pd.read_csv('adult.data')</pre>	

Retention

	RETENTION DURATION	RETENTION POLICY
	Specify the duration for which this dataset can be retained:	Summarize the retention policy for this dataset. Use this space to include any other information or links that might be relevant.
	<ul style="list-style-type: none">Not applicable- Publicly available data with no retention policy	<ul style="list-style-type: none">Not applicable- Publicly available data with no retention policy
	RETENTION STEPS	EXCEPTIONS AND EXEMPTIONS
	Summarize any additional requirements and related steps to retain the dataset.	Summarize any additional exceptions and related steps to retain the dataset:
	<ul style="list-style-type: none">Not applicable	<ul style="list-style-type: none">Not applicable

Wipeout and Deletion

	WIPEOUT DURATION	DELETION EVENT
	<i>Specify the duration after which this dataset should be deleted or wiped out:</i>	<i>Summarize the sequence of events and allowable processing for data deletion:</i>
	<ul style="list-style-type: none">Not applicable- Publicly available data with no wipeout and deletion policy.	<ul style="list-style-type: none">Not applicable
	ACCEPTABLE MEANS OF DELETION	POST-DELETION OBLIGATIONS
	<i>List the acceptable means of deletion:</i>	<i>Summarize the sequence of obligations after a deletion event:</i>
	<ul style="list-style-type: none">Not applicable	<ul style="list-style-type: none">Not applicable
	OPERATIONAL REQUIREMENTS	EXCEPTIONS AND EXEMPTIONS
	<i>List any wipeout integration operational requirements:</i>	<i>Summarize any additional exceptions and related steps to a deletion event:</i>
	<ul style="list-style-type: none">Not applicable	<ul style="list-style-type: none">Not applicable

Dataset Provenance

Data Collection & Sources

DATA COLLECTION METHODS	DATA SOURCES	DESCRIPTION OF DATA SOURCE(S)
<i>Bold to select all applicable. 👉 Do not delete any unselected choices.</i>	<i>Describe the source for each collection method. Add rows as meaningful. Refer to guidance on Duplicate for each collection method as necessary.</i>	<i>Provide a brief description of each Data Source by type. Include appropriate breakdowns if data sources contain data from other sources. Include links to more information, metrics, visualizations, etc.</i>
API Artificially Generated Crowdsourced - Paid Crowdsourced - Volunteer Vendor Collection Efforts Scraped or Crawled Survey, forms or polls Taken from other existing datasets Unknown To be determined Others (please specify)	Dataset was extracted from another public database 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)) The dataset was first published as part of a research article in 1996. [Source & Link]: https://archive.ics.uci.edu/ml/datasets/adult Date of Collection: 1994 Census database	1994 Census database Dataset was extracted from another public database 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))
DATASET TYPE	COLLECTED DATA	DATA PROCESSING
<i>Bold to select all applicable. 👉 Do not delete any unselected choices.</i>	<i>List or describe any fields or data that were collected for this dataset, and indicate if they were included in the dataset or excluded from the dataset. Include links, descriptive statistics, and visualizations where relevant. Duplicate for each collection method as necessary.</i>	<i>If multiple methods were used to collect data, how was the data aggregated, processed, or connected? Include relevant descriptions, statistics, metrics or visualizations, links and libraries in your response. Break down by source type.</i>
Static Data was collected once from single or multiple sources. Streamed Data is continuously acquired from single or multiple sources. Dynamic Data is updated regularly from single or multiple sources. Others* (*please specify)	Static Dataset was extracted from another database. <ul style="list-style-type: none">1994 Census Database	Extraction from 1994 Census Database A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))

Criteria		
SELECTION CRITERIA	INCLUSION CRITERIA	EXCLUSION CRITERIA
<i>Please describe the data selection criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant.</i>	<i>Please describe the data inclusion criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant..</i>	<i>Please describe the data exclusion criteria. Break down by method as applicable. Include links, descriptive statistics, and visualizations where relevant.</i>
<p>Extraction</p> <p>Data Extracted from 1994 Census database with some criteria.</p>	<p>Extraction</p> <p>Data extracted from 1994 Census database where age is greater than 16 , Adjusted gross income is greater than 100, final weight is greater than 1 and work hours per week is greater than 0.</p>	<p>Extraction</p> <ul style="list-style-type: none"> • Data extracted from 1994 Census database. • Some data were excluded where age is less than 16 , Adjusted gross income is less than 100, final weight is less than 1 and work hours per week is 0.
Relationship to Source		
USE	BENEFITS AND VALUE	LIMITATIONS AND TRADE-OFFS
<i>If at all, how is the resulting dataset aligned with the purposes, motivations, or intended use of the upstream source(s)? Break down by source type.</i>	<i>What are the benefits of the resulting dataset to its consumers, compared to the upstream source(s)? Break down by source type.</i>	<i>What are the limitations of the resulting dataset to its consumers, compared to the upstream source(s)? Break down by source type.</i>
<p>Census 1994 database included other data which were not useful for building a classification model. This dataset extracted useful data related to income so one can classify.</p> <p>Census 1994 Database</p>	<p>One can directly build a classification model for Income without cleaning and removing unnecessary data.</p>	<p>Not applicable</p>

Updates to Dataset

⚠ Fill this next row if: this is not the first version of the dataset, and there is no data card available for the first version.

	FIRST VERSION	NOTES ON FIRST VERSION
	Provide a basic description of the first version of this dataset.	Optional. Provide a short summary describing caveats or nuances of the first version of this dataset. Include links, charts, and visualizations as appropriate.
	Deprecated Current Version1.0 Last Updated05/1996 Release Date05/1996	Based on extraction from 1994 Census dataset. Visualization: and other details: https://epistasislab.github.io/pmlb/profile/adult.html
DATASET UPDATE FREQUENCY	DATASET UPDATE SCHEDULE	CHANGES ON UPDATE
Bold to select ONE 👉 Do not delete any unselected choices.	Please describe the update schedule	What happens when the dataset is refreshed? Break down by sources as necessary. Include any applicable policies and changes to the dataset that occur during a refresh.
Yearly Quarterly Monthly Biweekly Weekly Daily Hourly Static Others* (*Deprecated)	Deprecated The dataset is deprecated and no longer updated.	Deprecated and hence no longer maintained. However, is publicly available and is popular for classification models.

Human and Other Sensitive Attributes

SENSITIVE HUMAN ATTRIBUTES	INTENTIONALITY OF COLLECTIONS	RATIONALE FOR COLLECTING HUMAN ATTRIBUTES
<p><i>Bold to select ALL ATTRIBUTES that are present in the dataset.</i></p> <p>👉 <i>Do not delete any unselected choices.</i></p>	<p><i>For each human attribute indicated, specify if this information was collected intentionally or unintentionally:</i></p>	<p><i>Briefly describe the motivation, rationale, considerations or approaches that caused this dataset to include the indicated human attributes. Summarize why or how this might affect the use of the dataset.</i></p>
<p>Race</p> <p>Gender</p> <p>Ethnicity</p> <p>Socio-economic status</p> <p>Geography</p> <p>Language</p> <p>Sexual Orientation</p> <p>Religion</p> <p>Age</p> <p>Culture</p> <p>Disability</p> <p>Experience or Seniority</p> <p>Others (please specify)</p>	<p>Intentionally Collected Attributes (Human attributes that were labeled or collected as a part of the dataset creation process)</p> <ul style="list-style-type: none">● Race: Fields: Race● Gender: Fields: Gender● Age: Fields: Age● Ethnicity: Fields: Native-Country● Socio-economic Status: Fields: Education, Education num, martial-status, Occupation, relationship, capital-gain, capital-loss, work-class, hours-per-week.	<p>Dataset has all anonymous data which have sensitive human attributes, which was mainly collected from census data as these attributes are necessary to predict income.</p>
	SOURCE(S) OF HUMAN ATTRIBUTES	COLLECTION METHODS
	<p><i>Indicate the source of the sensitive attributes using the format provided.</i></p>	<p><i>Describe the methods used to collect human attributes in the dataset. Break down by human attribute as necessary. Include information related to the tasks, platforms, visualizations, links to additional documentation as applicable.</i></p>
	<p>[Human Attribute]: Extraction was done by Barry Becker from the 1994 Census database https://archive.ics.uci.edu/ml/datasets/adult</p>	<p>Extraction was done by Barry Becker from the 1994 Census database.</p>

DISTRIBUTION OF HUMAN ATTRIBUTES

Duplicate and populate the following row for each human attribute previously selected. Include the key takeaways in the caption.

Statistic	Age	Education-Num	Capital Gain	Capital Loss	Hours per week
count	45222	45222	45222	45222	45222
mean	38.547941	10.118460	1101.430344	88.595418	40.938017
std	13.217870	2.552881	7506.430084	404.956092	12.007508

min	17.00	1.00	0	0	1.00
25%	28.00	9.00	0	0	40.00
50%	37.00	10.00	0	0	40.00
75%	47.00	13.00	0	0	45.00
max	90.00	16.00	99999.00	4356.00	99.00
mode	36	9	0	0	40

Caption for table above

	KNOWN CORRELATIONS	RISK, TRADE-OFFS AND CAVEATS
	<i>List or describe any known correlations with the indicated sensitive attributes in this dataset. Summarize why or how this might affect the use of the dataset. Include visualizations, metrics, or links where necessary.</i>	<i>Provide a statement, list or summarize any expectations, systemic or residual risks, trade-offs and caveats due to human attributes in this dataset. Break down by human attribute if necessary.</i>
	Work class and occupation, Income and marital status, Income, and relationship	Not available


Extended Use		
Use with Other Data		
SAFETY OF USE WITH OTHER DATA	KNOWN SAFE DATASETS OR DATA TYPES	BEST PRACTICES FOR JOINING OR AGGREGATING WITH DATASET
<i>Bold to select ONE.</i> 👉 Do not delete any unselected choices.	<i>Which known datasets or data can this dataset be safely joined or aggregated with? Describe any relevant transformation types.</i>	<i>Summarize best practices for using this dataset in conjunction with other datasets or data type. Links to demonstrative examples where available.</i>
Safe to use with other data Conditionally safe to use with other data Should not be used with other data Unknown Others* (Please specify)	Dataset or Data Dataset is safe to use with other data without any restrictions.	Can use this dataset safely with other datasets provided they have same structure.
	KNOWN UNSAFE DATASETS OR DATA TYPES	KNOWN LIMITATIONS AND RECOMMENDATIONS
⚠️ Fill out this row if you selected “Conditionally safe to use with other datasets” or “Should not be used with other datasets”:	<i>Which known datasets or data should this dataset not be joined or aggregated with? List and describe any relevant transformation types.</i>	<i>Describe limitations of the dataset that might introduce foreseeable risks to intended use when the dataset is conjoined with other datasets. Include any suggested recommendations.</i>
	Not applicable	<ul style="list-style-type: none"> Dataset contains mostly Male and for race it’s mostly White so will not be useful for diverse problems.
Forking & Sampling		
SAFETY OF FORKING / SAMPLING	ACCEPTABLE SAMPLING METHODS	BEST PRACTICES FOR FORKING AND SAMPLING
<i>Bold to select ONE.</i> 👉 Do not delete any unselected choices.	<i>Bold to select all applicable.</i> 👉 Do not delete any unselected choices	<i>Summarize best practices for forking or sampling this dataset. Links to demonstrative examples where available.</i>
Safe to fork and/or sample Conditionally safe to fork and/or sample Should not be forked and/or sampled Unknown Others* (*Please specify)	Cluster Sampling Haphazard Sampling Multi-stage Sampling Random Sampling Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others* (*Please Specify)	If training model to predict accuracy for income prediction the overall accuracy is different and when looking into accuracy depending on gender there is a performance gap, hence sampling data with gender is a good idea to reduce performance gap.
	KNOWN RISKS TO SAMPLING	KNOWN LIMITATIONS AND RECOMMENDATIONS


⚠ Fill out this row if you selected “Conditionally safe to fork and/or sample” or “Should not be forked and/or sampled”.	What known or residual risks are associated with forking and sampling methods when applied to the dataset? List and describe.	Describe limitations of the dataset that might introduce foreseeable risks to intended use when the dataset is forked or sampled. Include any suggested recommendations.
	<ul style="list-style-type: none"> NA 	<ul style="list-style-type: none"> If training model to predict accuracy for income prediction the overall accuracy and accuracy depending on gender may differ, hence sampling data with gender is a good idea to reduce performance gap.

Use in Machine Learning or AI Systems

DATASET USE(S)	DATASET SPLITS	USAGE GUIDELINES OR POLICIES						
<p><i>Bold to select all applicable.</i></p> <p>👉 Do not delete any unselected choices.</p>	<p>Describe and name the splits in the dataset (if more than one), and include any criteria for splitting the data.</p>	<p>Describe any usage guidelines or policies that users of the dataset should be aware of. Summarize documents and link to them as relevant.</p>						
<div> <div>Training</div> <div>Testing</div> <div>Validation</div> <div>Dev</div> <div>Others*</div> <div>(* Please Specify)</div> </div>	<div> <div>Train</div> <div>32561</div> <div>Test</div> <div>16281</div> </div>	<div>Not applicable</div>						
	FEATURE DISTRIBUTIONS	KNOWN CORRELATIONS						
	<p>Describe any notable feature distributions in the dataset. Include links to servers where readers can explore the data on their own.</p>	<p>List or describe any known correlations with the indicated features in this dataset. Summarize why or how this might affect the use of the dataset. Include links where necessary.</p>						
	<p>Feature and descriptive statistics for the data including visualization:</p> <p>https://epistasislab.github.io/pmlb/profile/adult.html</p>	<p>Work class and occupation, Income and marital status, Income, and relationship. These features are correlated because they have direct influence on the income, as it makes sense person with higher income is likely to be married vs someone with no or less income.</p>						
	SPLIT STATISTICS							
	<p>Provide the sizes of each split. As appropriate, provide any descriptive statistics for features.</p>							
	<table> <tr> <td>Statistic</td><td>Train</td><td>Test</td></tr> <tr> <td>Count</td><td>32561</td><td>16281</td></tr> </table> <p>Caption for table above</p>		Statistic	Train	Test	Count	32561	16281
Statistic	Train	Test						
Count	32561	16281						

Dataset Transformations

 Fill this section if any transformations were applied in the creation of your dataset.

TRANSFORMATIONS APPLIED	FIELDS TRANSFORMED	LIBRARIES AND METHODS USED
<i>Bold to select all applicable</i>  Do not delete any unselected choices.	<i>What were the data types that fields were transformed to? Break down by transformations applied</i>	<i>List any relevant libraries used to process the data, as applicable.</i>
Anomaly Detection Cleaning Mismatched Values Cleaning Missing Values Converting Data Types Data Aggregation Dimensionality Reduction Joining Input Sources Redaction or Anonymization Others* (*No Transformation)	Not applicable	<ul style="list-style-type: none">Not applicable

Breakdown of Transformations

Fill out relevant rows.

CLEANING MISSING VALUES	METHODS USED	COMPARATIVE SUMMARY
<i>Which fields in the data were missing values? How many?</i>	<i>How were missing values cleaned? What other choices were considered?</i>	<i>Why were missing values cleaned using this method (over others)? Provide comparative charts showing before and after missing values were cleaned.</i>
Not applicable	< Not applicable	Not applicable
CLEANING MISMATCHED VALUES	METHODS USED	COMPARATIVE SUMMARY
<i>Which fields in the data were corrected for mismatched values?</i>	<i>How were incorrect or mismatched values cleaned? What other choices were considered?</i>	<i>Why were incorrect or mismatched values cleaned using this method (over others)? Provide a comparative analysis demonstrating before and after values were cleaned.</i>
Not applicable	Not applicable	Not applicable
ANOMALY DETECTION	METHODS USED	OUTLIERS HANDLING
<i>How many anomalies or outliers were detected?</i>	<i>What methods were used to detect anomalies or outliers?</i>	<i>If at all, how were anomalies or outliers handled? Why or why not?</i>
Not applicable	Not applicable	Not applicable

DATA AGGREGATION	METHODS USED	COMPARATIVE SUMMARY
Which fields in the dataset were aggregated?	What methods were used to aggregate the data? Include the aggregating operator. What other choices were considered?	Why was the data aggregated using this method (over others)? Provide comparative charts that demonstrate the choices of aggregators.
Not applicable	Not applicable	Not applicable
DIMENSIONALITY REDUCTION	METHODS USED	COMPARATIVE SUMMARY
How many original features were collected and how many dimensions were reduced?	What methods were used to reduce the dimensionality of the data? What other choices were considered?	Why were features reduced using this method (over others)? Provide comparative charts showing before and after dimensionality reduction processes.
Not applicable	Not applicable	Not applicable
JOINING INPUT SOURCES	METHODS USED	RESIDUAL RISKS AND APPROVALS
What were the distinct input sources that were joined?	What are the shared columns of fields used to join these sources?	What are the differential privacy or other residual risks from this join? Include links to relevant approvals and documentation.
Not applicable	Not applicable	Not applicable
REDACTION OR ANONYMIZATION	METHODS USED	RESIDUAL RISKS AND APPROVALS
Which features were redacted or anonymized?	What methods were used to redact or anonymize data?	What are the differential privacy or reidentification risks to redacted data or anonymization? Include links to relevant approvals and documentation.
Not applicable	Not applicable	Not applicable
OTHERS (PLEASE SPECIFY)	METHODS USED	RESIDUAL RISKS & COMPARATIVE SUMMARY
What was done? Which features or fields were affected?	What methods were used?	What are the residual risks associated with this transformation? Include links to relevant approvals and documentation. Why were features reduced using this method (over others)? Provide comparative charts showing before and after this transformation.
Not applicable	Not applicable	Not applicable

Annotations

⚠ Fill this section if any human or algorithmic annotation tasks were performed in the creation of your dataset.

ANNOTATION WORKFORCE TYPE	ANNOTATION CHARACTERISTICS	ANNOTATION DESCRIPTION
<i>Bold to select ALL APPLICABLE</i> 👉 Do not delete any unselected choices.	<i>Describe relevant characteristics as indicated. For quality metrics, consider including accuracy, consensus accuracy, IRR, XRR at the appropriate granularity (e.g. across dataset, by annotator, by annotation, etc.). Duplicate for each annotation type if multiple methods were used.</i>	<i>Briefly describe the annotations applied to the dataset, including but not limited to: Creation of data, authoring of data, labeling, annotation, rating, etc. Include links, and indicate platforms, tools or libraries used wherever possible. Break down by annotation type as applicable.</i>
Annotation Target in Data Machine-generated Annotations Human Annotations - Expert Human Annotations - Non-expert Human Annotations - Employees Human Annotations - Contractors Human Annotations - Crowdsourcing Human Annotations - Outsourced / Managed Teams Unlabeled Others* (*Please specify)	The data was extracted from 1994 Census database. 1994 Census database was gathered through human annotations.	Not applicable <ul style="list-style-type: none">Data extracted from another database.
	ANNOTATION DISTRIBUTION(S)	ANNOTATION TASK AND INSTRUCTIONS
	<i>Provide a distribution of annotations for each annotation or class of annotations using the format below. Duplicate for each annotation type if multiple methods were used.</i>	<i>Briefly summarize the annotation task and instructions provided to annotators or methods employed for machine annotations. Include the inter-annotation adjudication policy, and any golden questions if applicable. Add links wherever possible. Break down by annotation type as applicable.</i>
	Not applicable	Not applicable

Description of Human Annotators

⚠ Fill this section if human annotators were used.

	ANNOTATOR BREAKDOWN	ANNOTATOR DESCRIPTION
	Provide a description of the annotators. Add more rows as meaningful. For inapplicable rows, refer to guidance on slide 38 of go/recommended-by. Duplicate for each annotation type if multiple methods were used.	Provide a brief description of the annotator pool(s). Elaborate on the annotator type, training provided, selection criteria, and anything else that might affect the quality of annotations. Break down by annotation type.
	Not applicable	Not applicable
LANGUAGE(S) OF ANNOTATORS	LOCATION(S) OF ANNOTATORS	GENDER(S) OF ANNOTATORS
Provide distributions as available. Duplicate for each annotation type if multiple methods were used.	Provide distributions as available. Duplicate for each annotation type if multiple methods were used.	Provide distributions as available. Duplicate for each annotation type if multiple methods were used.
Not applicable	Not applicable	Not applicable

Validation Methods		
<div>⚠ Fill this section if the data in dataset was validated during or after the creation of your dataset.</div>		
Validation Method(s)	Validation Breakdown	Description of Validation
<div>Bold to select ALL APPLICABLE</div> <div>👉 Do not delete any unselected choices.</div>	<div>Describe the fields and data points that were validated. Duplicate for each validation type if multiple methods were used.</div>	<div>Briefly describe the methods used to validate the dataset. Include tools, frameworks, libraries, platforms used. Indicate results, outcomes, actions and visualizations. Include links wherever possible. Break down by validation type if multiple methods were used.</div>
Data Type Validation Range and Constraint Validation Code/cross-reference Validation Structured Validation Consistency Validation Not Validated Others* (*Please specify)	Not applicable	Not applicable
Description of Human Validators		
<div>⚠ Fill this section if the dataset was validated using human validators</div>		
	Validators Characteristic(s)	Validators Description(s)
	<div>Describe the following about the validators. Add more rows as meaningful. Duplicate for each validation type if multiple methods as necessary.</div>	<div>Provide a brief description of each validator pool. Elaborate on the annotator type, training provided, selection criteria, and anything else that might affect the quality of annotations. Break down by validation type as necessary.</div>
	Not applicable	Not applicable
Language(s) of Validators	Location(s) of Validators	Gender(s) of Validators
<div>Provide the following distribution as available. Duplicate for each validation type as necessary.</div>	<div>Provide the following distribution as available. Duplicate for each validation type as necessary.</div>	<div>Provide the following distribution as available. Duplicate for each validation type as necessary.</div>
Not applicable	Not applicable	Not applicable

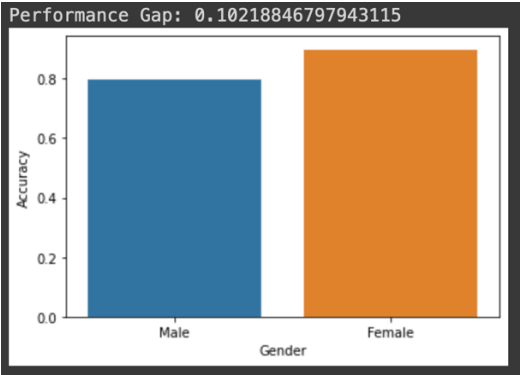
Sampling Methods

⚠ Fill out the following block if your dataset employed any sampling methods.

SAMPLING METHOD(S)	SAMPLING CHARACTERISTIC(S)	SAMPLING CRITERIA
<i>Bold to select ALL APPLICABLE</i> <i>👉 Do not delete any unselected choices.</i>	<i>Provide the following for each sampling method used. Add additional sampling statistics as relevant. Duplicate for each sampling type, if multiple methods were used.</i>	<i>Describe any criteria used to sample the data. Break down by sampling methods as relevant. Include links and metrics where necessary.</i>
Cluster Sampling Haphazard Sampling Multi-stage Sampling Random Sampling Retrospective Sampling Stratified Sampling Systematic Sampling Weighted Sampling Unknown Unsampled Others* (*Extraction)	Data extracted from 1994 Census Database.	Data was extracted from 1994 Census Database.


Known Applications & Benchmarks

⚠ Fill out the following section if your dataset was primarily created for use in AI or ML system(s)

ML APPLICATION(S)	EVALUATION - RESULTS	EVALUATION - PROCESS
<p>✎ Write tags separated by commas. Focus on key tasks performed by the model</p>	<p>Enumerate the models on which this dataset was used and corresponding performance metrics. Link to model cards or model documentation. Duplicate for each model.</p>	<p>Describe any notable factors in your process for evaluating your model's overall performance or assessing how the dataset contributes to the model's performance.</p> <p>Break down for each model. Include links, metrics, charts, and visualizations.</p>
<p>Classification, Regression</p>	<p>https://colab.research.google.com/drive/1mPLMcvalmGErsZO8-BhSUDI03hWNV6u?usp=sharing</p> <p>Accuracy 79.58%</p>	<p>Overall accuracy is around 79.58% when it comes to income prediction with just one layer of neuron network.</p>
	<p>MODEL DESCRIPTION(S) AND STATISTICS</p> <p>Performance Gap: 0.10218846797943115</p> 	<p>EXPECTED PERFORMANCE AND KNOWN CAVEATS</p> <p>Overall accuracy is around 79.58% however there is a difference in performance when it comes to gender there is a high-performance gap.</p>
<p>Bold to select ONE</p> <p>✎ Do not delete any unselected choices.</p>	<p>Briefly describe the model(s) and tasks that this dataset was used in. Include links where necessary. Duplicate for each model.</p>	<p><i>Expected performance:</i> Briefly summarize the application and expected performance when using this dataset.</p> <p><i>Known Caveats:</i> Describe the known caveats, trade-offs and consequences when using this dataset. Duplicate for each model. Include links wherever possible.</p>
<p>Duplicate this row as necessary for each model type</p>	<p>Used neuron network with one unit to train and predict income.</p>	<p>Used neuron network with one unit to train and predict income.</p> <ul style="list-style-type: none"> Overall accuracy is around 79.58% however there is a difference in performance when it comes to gender there is a high-performance gap.

Terms of Art

Concepts and Definitions referenced in this Data Card


 Use this space to include the expansions and definitions of any acronyms, concepts, or terms of art used across the Data Card. Use standard definitions where possible. Include the source of the definition where indicated. If you are using an interpretation, adaptation, or modification of the standard definition for the purposes of your data card or dataset, include your interpretation as well.

Fnlwgt

Definition: Final weight, “In other words, this is the number of people the census believes the entry represents”

Source: <https://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/048.pdf>

Reflections on Data

 Use this space to include any additional information about the dataset that has not been captured by the Data Card. For example, Does the dataset contain data that might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please contact the appropriate parties to mitigate any risks.

Personal Reflection

This is a popular dataset for classification machine learning. People use this in various models however, the data is not very diverse majority of the data have their gender as male, and are white, hence there is a performance gap when it comes to accuracy for different genders.

