Dataset Overview: Adult (Census Income)

Felipe S. Dosso, Betania E. R. da Silva, Décio M. Filho

September 10, 2024

Introduction

This dataset, also known as the "Census Income" dataset, is used for classification tasks in machine learning. It contains demographic and income-related information collected from the U.S. Census Bureau in 1996 by Barry Becker. The objective of this dataset is to predict whether a person earns over $50,000 per year based on their demographic attributes such as age, education, occupation, and work hours. The dataset has been widely used in machine learning benchmarks, particularly for binary classification models. The *Adult* datasets contains data in a tabular form and is in the raw space of data.

Dataset Overview

Characteristic	**Details**
Number of Instances	48,842
Number of Features	14 (Including categorical and continuous)
Classes	$\leq$ 50K, $>$ 50K
Sensitive Data	Race, Gender, Native Country

Table 1: Summary of Dataset Characteristics

Contact Information

- **UCI Machine Learning Repository**: https://archive.ics.uci.edu
- **Contact**: ml-repository@ics.uci.edu

Descriptive Statistics

Numerical Features

Statistic	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	30162	30162	30162	30162	30162	30162
mean	38.43	189793	10.12	1092	88.37	40.93
std	13.13	105630	2.54	7406	404.29	11.97
min	17	13769	1	0	0	1
max	90	1484705	16	99999	4356	99

Table 2: Descriptive Statistics for Numerical Attributes

Categorical features

Statistic	workclass	education	marital-status	occupation	relationship	race	sex	native-country
count	30162	30162	30162	30162	30162	30162	30162	30162
unique	7	16	7	14	6	5	2	41
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States
freq	22286	9840	14065	4038	12463	25932	20380	27504

Table 3: Descriptive statistics of categorical variables

Target

Statistic	income
count	30162
unique	2
top	$\leq$ 50K
freq	22654

Table 4: Descriptive statistics of target

Data Transformations

To prepare the data, transformations were made. The missing values were removed and the numerical attributes were normalized. Considering the labeling of sensitive variables, the protected attribute is labeled as '1' to facilitate the detection of inequality or bias. It was decided to assign the label '1' to the protected groups (Female, in the case of Gender; 'Non-White' in the case of Race;Immigrant, in the case of Country of Origin), while the label '0' is assigned to the non-protected group (Male, in the case of Gender; 'White' in the case of Race; Born in the USA, in the case of Country of Origin). After the binary categorization for the mentioned features, One-Hot encoding was applied to the remaining categorical features, resulting in the dataset now having 55 columns (instead of the initial 15).

Bias

Bias in machine learning can lead to unfair, inaccurate, or harmful outcomes, especially when decisions impact people's lives. When models are trained on data that reflects existing societal biases, they can unintentionally perpetuate and even amplify these inequalities.

Identifying and addressing bias is crucial for ensuring that machine learning systems are fair, transparent, and trustworthy. By detecting biases early, we can mitigate their negative effects.

Three types of bias are observed in this dataset, social bias, selection bias and temporal bias. Social and selection bias are very similar. Social bias occur when there is a under representation in marginalized groups. In the Adult dataset, more than 7,508 individuals have a high income (>50K), but only 513 of them were immigrants. This could lead to a model that favors americans for they are the privileged group.

While selection bias is when there is an imbalance between two or more groups in the same variable. For instance, there are 27,504 americans, but the number of non-americans is relatively low (2,658). This may lead to a model that performs poorly for immigrants.

Another existing bias is the difference of time between the data collection and the modeling of it. The Adult dataset was collected from the U.S. Census Bureau in 1996, but it's being used to model a machine learning system in 2024. This distance is problematic because the distribution of data in the past is probably different from the distribution in the present.

Although no method was used to mitigate the existing bias in the data, they were identified and the next steps will consider them.

Privacy

Data privacy is crucial when building machine learning systems, as it protects individuals' sensitive information from unauthorized access and misuse. When personal data is used without proper safeguards, it can lead to identity theft, discrimination, or breaches of confidentiality. Ensuring privacy in data not only complies with regulations, but also fosters ethical practices. By safeguarding privacy, we create machine learning models that respect individuals' rights while maintaining the integrity and security of the data used in training. Among the methods to ensure privacy, two procedures stand out, k-anonymity and differential privacy.

k-anonymity is a key concept in data privacy that aims to protect individuals' identities in datasets. It achieves this by ensuring that each record is indistinguishable from at least $k - 1$ other records with respect to certain identifying attributes. This means that for any combination of quasi-identifiers (attributes that can potentially identify an individual, such as ZIP code, age, and gender), there will be at least k records that share the same values, making it difficult to pinpoint any single individual.

To achieve k-anonymity, data can be modified through techniques like generalization and suppression. Generalization involves replacing specific values with broader categories (e.g., replacing exact ages with age ranges), while suppression involves removing or hiding some data points altogether. These methods help obscure individual identities while retaining useful information in the dataset.

Another important concept is differential privacy. At it's core, it's about ensuring that the outcomes of data queries remain nearly the same, regardless of whether any specific individual's data is included. This concept is quantified using a parameter known as ϵ. A lower ϵ indicates stronger privacy protection, while a higher means the privacy guarantees are weaker. Essentially, differential privacy seeks to mask the influence of any single individual's data on the final results.

To achieve differential privacy, noise is added to the output of data queries. This noise is generated from a statistical distribution, such as the Laplace distribution, and serves to obscure the contribution of any single individual. The amount of noise introduced depends on the sensitivity of the query and the desired level of privacy.

Another crucial aspect of differential privacy is the management of the privacy parameter ϵ. This parameter determines the trade-off between privacy and data accuracy. While a smaller ϵ offers better privacy protection, it can result in less accurate results. Conversely, a larger ϵ provides more precise results but with reduced privacy assurances.

Moreover, differential privacy can be maintained across multiple queries or analyses by carefully controlling how the privacy budget is allocated. This ensures that privacy is preserved even when multiple pieces of information are extracted from the dataset.

In conclusion, data privacy is a fundamental concern in today's data-driven world. As organizations increasingly rely on vast amounts of personal and sensitive information, safeguarding this data against unauthorized access and misuse becomes paramount. Effective data privacy practices, such as employing techniques like k-anonymity and differential privacy, help ensure that individual identities remain protected while still enabling meaningful data analysis.

Data Dictionary

Attribute	**Type**	**Description**
age	Integer	Age of the individual
workclass	Categorical	Type of employment
fnlwgt	Integer	Final weight for sampling
education	Categorical	Highest education level
marital-status	Categorical	Marital status
occupation	Categorical	Type of job
relationship	Categorical	Family relationship
race	Categorical	Race
sex	Categorical	Gender
capital-gain	Integer	Capital gains
capital-loss	Integer	Capital losses
hours-per-week	Integer	Work hours per week
income	Categorical	Income bracket ($\leq$50K, >50K)

Table 5: Dataset Attributes