HW1

Abstract

K-Anonymity has several techniques such as attribute suppression, data masking, and data de-identification. We will proceed according to the process of anonymization: first, understand the data; second, use anonymization techniques; third, evaluate de-identification; fourth, apply K-Anonymity to describe our work.

Introduction

In this homework, we need to implement k-anonymity on the Adult dataset which is from UC Irvine. The task needs to predict whether annual income of an individual exceeds $50K/yr based on census data, also known as the 'Census Income' dataset. This dataset has categorical and integer feature types, and it has fourteen features. For k-anonymity, we refer to optimization-based k-anonymity algorithms from Liang et al.'s work, and 'A Comparison of the Effects of K-Anonymity on Machine Learning Algorithms' by Wimmer et al.

Related Work

Privacy-Preserving Data Mining and Machine Learning Performance

Previous research in Privacy-Preserving Data Mining (PPDM) has primarily focused on how to protect data privacy through various anonymization techniques, without comprehensive analysis of their impact across different machine learning models. Wimmer et al. conducted a comparative study on the effects of K-anonymity on various machine learning algorithms. Their work demonstrated that different models respond differently to anonymized data, with C4.5 Decision Trees showing the best performance, followed by Naive Bayes and Logistic Regression, while Artificial Neural Networks, Decision Stump, and CART exhibited poorer performance. This research is significant as it addresses the inherent trade-off between privacy protection and model accuracy, providing insights into which algorithms maintain better predictive power when working with anonymized datasets.

K-Anonymity Techniques and Implementation Process

K-anonymity has emerged as a foundational privacy-preserving technique in data mining. Liang et al. explored optimization-based K-anonymity algorithms that balance privacy protection with data utility. Their work, along with other research, has established several common techniques for achieving K-anonymity, including attribute suppression (removing sensitive identifiers), record suppression (deleting specific records that cannot be properly anonymized), data masking (replacing or altering portions of data), pseudonymization (using randomly generated codes to replace personal identifiers), and generalization (converting specific values into broader categories). The implementation process of K-anonymity typically follows four key steps: understanding the data's characteristics, applying appropriate anonymization techniques, evaluating de-identification risk, and finally implementing K-anonymity to ensure that each record is indistinguishable from at least K-1 other records, thereby preventing re-identification attacks while preserving valuable information for analysis.

Method

Data Preprocessing

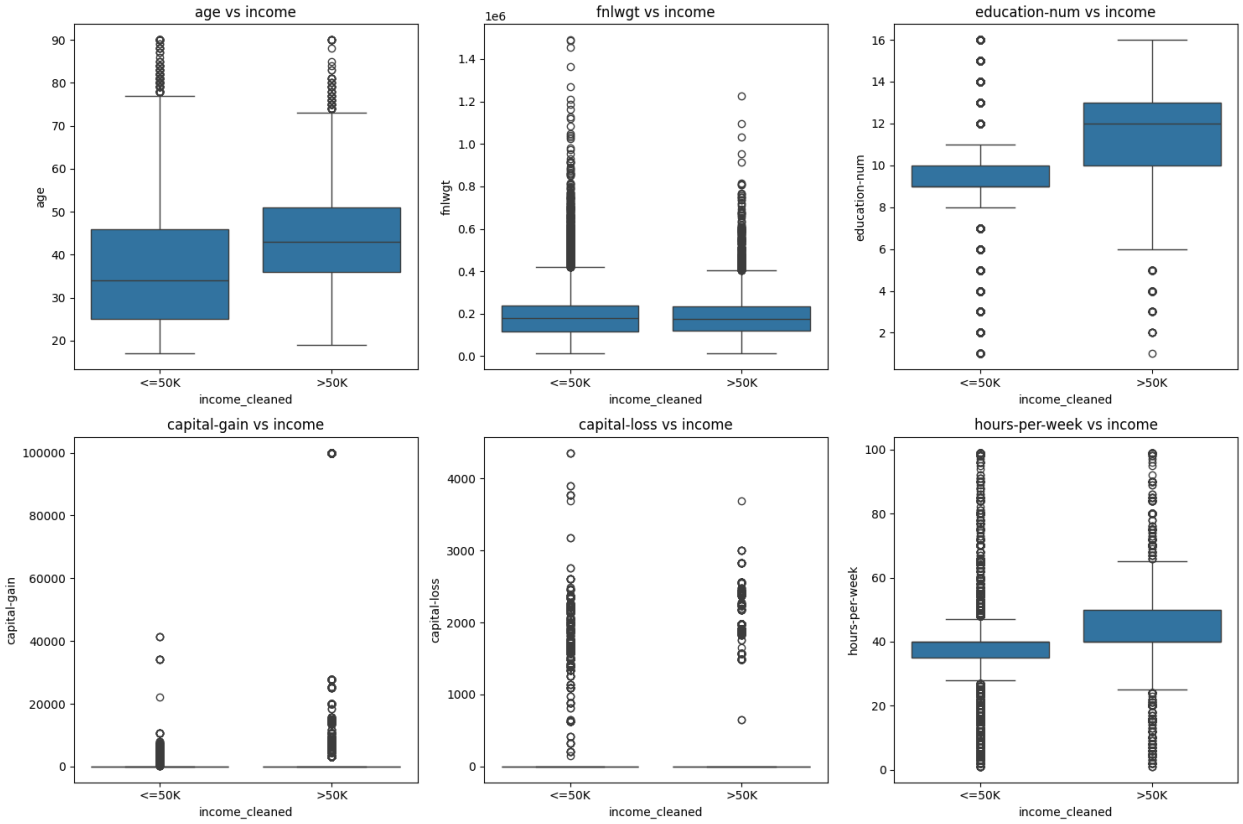
Clean null value and wrong format

First of all, we clean the income label because some values have a period at the end while others do not. For example, in the original data, we have 24,720 records labeled as '<=50K' and 12,435 records labeled as '<=50K.' with a period. We merge these into a single category, resulting in 37,155 records labeled as '<=50K' in our cleaned dataset. Similarly, we combine the '>50K' and '>50K.' categories into one '>50K' group with 11,687 records. Then, for the 'workclass', 'occupation', and 'native-country' columns that contain null values and question marks, we fill these missing data with the mode value of each respective column.

Investigate numeric features

We examined numeric features in relation to income classes by calculating the mean values of each numeric column grouped by income. We then created boxplots to visualize the distribution of each numeric feature across income

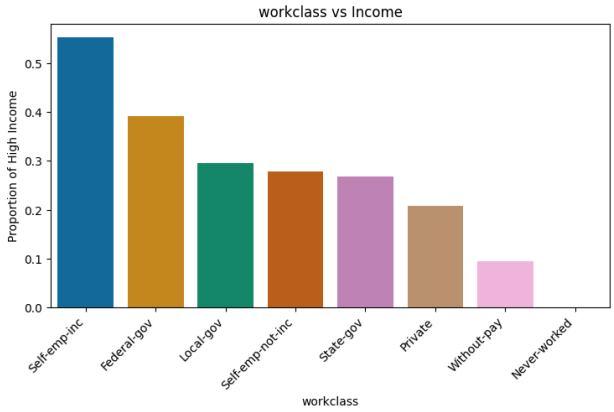
groups. We found that `fnlwgt` cannot effectively distinguish between income categories (over or under 50K). Both `capital-loss` and `capital-gain` require preprocessing or grouping due to their skewed distributions with many outliers. We also investigated `age`, `education-num`, and `hours-per-week`, which may contain outliers but appear more useful than the other three features for classifying whether income exceeds fifty thousand dollars. The boxplots clearly show that individuals with higher education levels, more working hours per week, and greater age tend to have higher incomes.

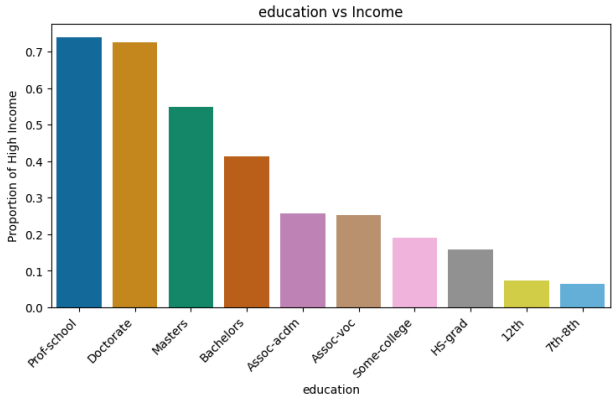


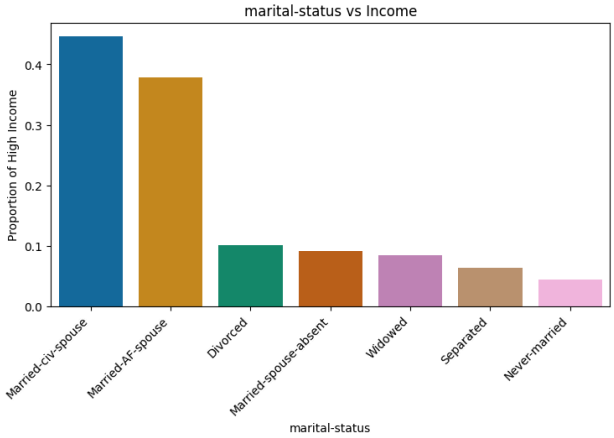
Investigate categorical features

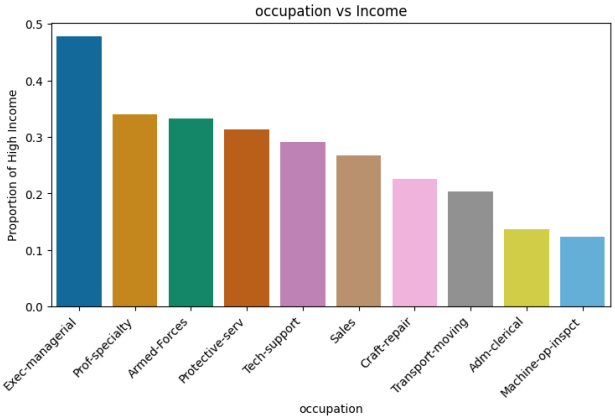
For the categorical features, we created bar plots showing the proportion of high income (>$50K) within each category. We used crosstabs to calculate the percentage of high income earners across different categorical values and sorted them to identify the most significant patterns.Based on the visualizations, we observed that `education`, `occupation`, and `workclass` appear to be strong predictors for distinguishing between income levels above or below $50K. The charts clearly show that higher education levels (Professional, Doctorate, Masters) have much

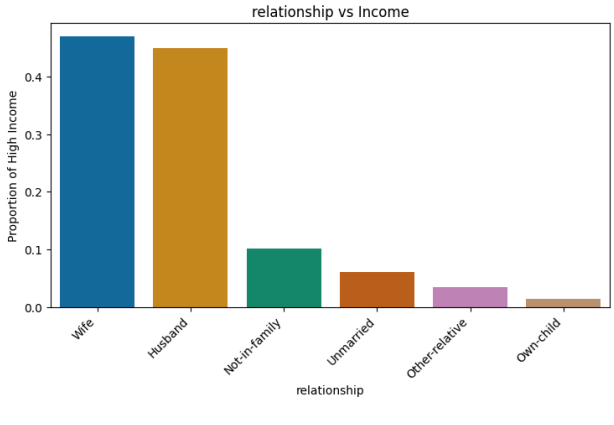
higher proportions of high income compared to lower education levels. Similarly, certain occupations (Executive-managerial) and workclass categories (Self-employed) show stronger associations with higher income. Other notable patterns include marital status, where `Married-civ-spouse` has a much higher proportion of high income than other categories, and sex, which shows a significant disparity between males and females. The relationship category also shows 'Wife' and 'Husband' with notably higher income proportions than other relationship statuses.

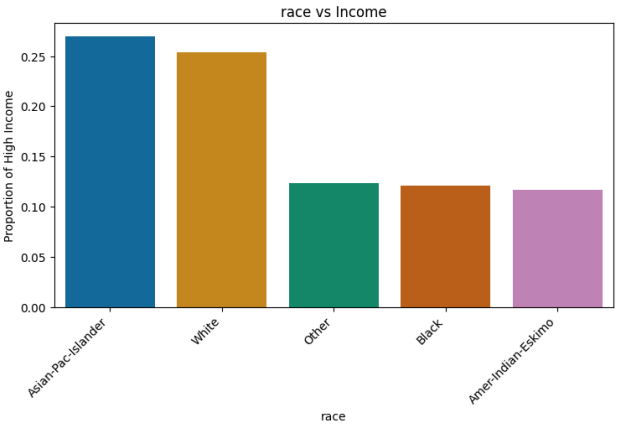

workclass vs Income
Proportion of High Income
0.0
0.1
0.2
0.3
0.4
0.5
Self-emp-inc
Federal-gov
Local-gov
Self-emp-not-inc
State-gov
Private
Without-pay
Never-worked
workclass

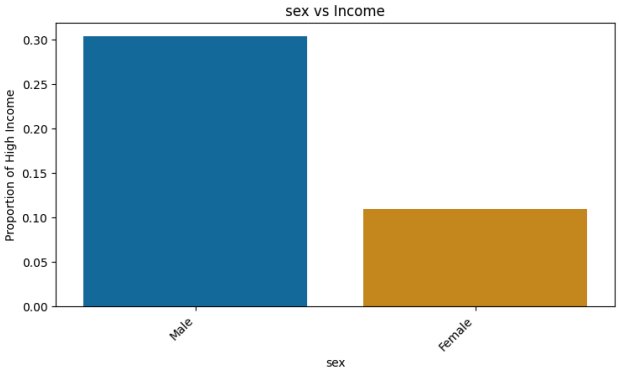

education vs Income
Proportion of High Income
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Prof-school
Doctorate
Masters
Bachelors
Assoc-acdm
Assoc-voc
Some-college
HS-grad
12th
7th-8th
education

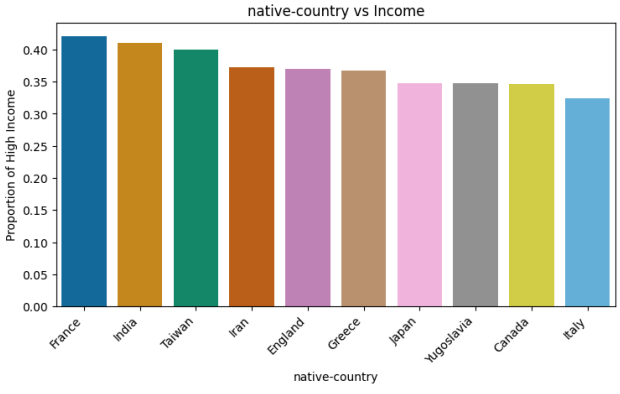

marital-status vs Income
Proportion of High Income
0.0
0.1
0.2
0.3
0.4
Married-civ-spouse
Married-AF-spouse
Divorced
Married-spouse-absent
Widowed
Separated
Never-married
marital-status


occupation vs Income
Proportion of High Income
0.0
0.1
0.2
0.3
0.4
0.5
Exec-managerial
Prof-specialty
Armed-Forces
Protective-serv
Tech-support
Sales
Craft-repair
Transport-moving
Adm-clerical
Machine-op-inspct
occupation


relationship vs Income
Proportion of High Income
0.0
0.1
0.2
0.3
0.4
Wife
Husband
Not-in-family
Unmarried
Other-relative
Own-child
relationship


race vs Income
Proportion of High Income
0.00
0.05
0.10
0.15
0.20
0.25
Asian-Pac-Islander
White
Other
Black
Amer-Indian-Eskimo
race


sex vs Income
Proportion of High Income
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Male
Female
sex


native-country vs Income
Proportion of High Income
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
France
India
Taiwan
Iran
England
Greece
Japan
Yugoslavia
Canada
Italy
native-country

Attribute Supression

Attribute Suppression will directly remove sensitive attributes from datasets to prevent personal data exposure. This includes eliminating uniquely identifying fields like "name," "ID number," and "address," as well as potentially removing other high-risk attributes that could contribute to re-identification when combined with other data. Suppression can be applied to entire columns or selectively to specific values within columns that pose significant privacy risks.

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income	income_cleaned
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K	<=50K

In this section, we check if the data shows any privacy information. As we can see, we don't need to remove any sensitive attributes from the dataset.

Pertub Methods and Privacy Level

We implemented a k-anonymity algorithm using the Split and Carry approach to protect privacy in datasets. The algorithm begins by identifying numeric columns and computing their variances. It then sorts data based on these variances and processes the dataset in chunks, creating equivalence classes of at least k records. For numeric attributes, the code generalizes values into ranges (min-max) to ensure individuals can't be uniquely identified. The implementation manages information loss by calculating how much precision is sacrificed during generalization. It also includes a carry-over mechanism to maintain equivalence classes across chunk boundaries. In preprocessing, we removed unnecessary features like `fnlwgt`, `capital-gain`, and `capital-loss`. Finally, we applied **`k=5`** anonymity with a subproblem size factor of **`S=4`**, balancing privacy protection with data utility. After k-anonymity, the dataset looks like the image below.

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	hours-per-week	native-country	income	income_cleaned	capital_ratio
0	[17.0-17.0]	Private	10th	[6.0-8.0]	Never-married	Other-service	Own-child	White	Female	[20.0-40.0]	United-States	<=50K.	<=50K	2174.0
1	[17.0-17.0]	Private	12th	[6.0-8.0]	Never-married	Handlers-cleaners	Own-child	White	Male	[20.0-40.0]	United-States	<=50K.	<=50K	0.0
2	[17.0-17.0]	Private	10th	[6.0-8.0]	Never-married	Prof-specialty	Own-child	White	Male	[20.0-40.0]	United-States	<=50K	<=50K	0.0
3	[17.0-17.0]	Private	11th	[6.0-8.0]	Never-married	Handlers-cleaners	Own-child	White	Male	[20.0-40.0]	United-States	<=50K.	<=50K	0.0
4	[17.0-17.0]	Never-worked	11th	[6.0-8.0]	Never-married	Prof-specialty	Own-child	Black	Female	[20.0-40.0]	United-States	<=50K.	<=50K	0.0

Dataset

This dataset is the UCI Adult Census Income dataset containing demographic and employment information about individuals with features such as age, workclass, education, marital status, occupation, race, sex, capital gains/losses, work hours, and native country, with the target variable being whether a person earns more or less than $50,000 annually. The data has been split into a training set (80%, 39,073 samples) and testing set (20%, 9,769 samples), both containing 9 selected features from the original dataset, making this a binary classification problem that requires appropriate preprocessing for its mix of categorical and numerical variables.

ML Model and Parameters

Logistic Regression

Metric	Value
Accuracy	0.7467
Precision	0.6744
Recall	0.7467
F1 Score	0.6775

Random Foreast

Parameters:

- n_estimators=100
- max_depth=15
- min_samples_split=5
- min_samples_leaf=2

Metric	Value
Accuracy	0.8343
Precision	0.8265
Recall	0.8343
F1 Score	0.8283

Support Vector Machine

Metric	Value
Accuracy	0.8015
Precision	0.7928
Recall	0.8015
F1 Score	0.7959

XGBoost

Parameters:

- random_state=seed
- n_estimators=200
- max_depth=5
- learning_rate=0.1

Metric	Value
Accuracy	0.8456
Precision	0.8391
Recall	0.8456
F1 Score	0.8409

References

Anonymity Techniques：https://zhuanlan.zhihu.com/p/652090236

Optimization-based *k*-anonymity algorithms:

https://www.sciencedirect.com/science/article/pii/S0167404820300377?fr=RR-2&ref=pdf_download&rr=920006996cdfd4cc

Kaggle Adult：https://www.kaggle.com/code/prashant111/eda-logistic-regression-pca

Paper: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=39e9e2b1f428b111d8efdab0422c981291287087

Dataset: https://archive.ics.uci.edu/dataset/2/adult