National Institute of Technology

Warangal

Lab 2 Assignment

Name: Ayush Rana

Roll Number: 24CSM2R05

Course Title: Data Privacy

Department: Computer Science and Engineering

Program: M. Tech in Computer Science and Information Security

Semester: 2

1. Implement (X,Y) -Anonymity for the adult data set via code.

[1]:
```
import pandas as pd
import numpy as np
from sklearn.preprocessing import KBinsDiscretizer

# Step 1: Load the dataset
columns = [
    "age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
    "occupation", "relationship", "race", "sex", "capital_gain",
    "capital_loss", "hours_per_week", "native_country", "income"
]

data = pd.read_csv("adult.data", names=columns, sep=r',\s*', engine='python', na_values='?')
```

[2]:
```
data
```

[2]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	

[2]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-

32561 rows × 15 columns

[3]:
```
# Step 2: Define quasi-identifiers (X) and sensitive attributes (Y)
quasi_identifiers = ["age", "education", "marital_status", "race", "sex"]
sensitive_attributes = ["income"]
```

[4]:
```
# Step 3: Generalize or bin quasi-identifiers (example: binning 'age')
def generalize_age(data, bins=10):
    binner = KBinsDiscretizer(n_bins=bins, encode='ordinal', strategy='uniform')
    data['age'] = binner.fit_transform(data[['age']]).astype(int)
    return data

data = generalize_age(data)
```

[5]:
```
# Step 4: Apply k-Anonymity by grouping and suppressing low-frequency combinations
k = 5  # Define the desired level of anonymity

def apply_k_anonymity(data, quasi_identifiers, k):
    # Group by quasi-identifiers
    grouped = data.groupby(quasi_identifiers)

    # Filter groups with fewer than k entries
    valid_groups = grouped.filter(lambda x: len(x) >= k)

    # Suppress small groups by replacing with "Suppressed"
    suppressed = grouped.filter(lambda x: len(x) < k).copy()
    for qi in quasi_identifiers:
        suppressed[qi] = "Suppressed"

    # Combine valid and suppressed groups
    anonymized_data = pd.concat([valid_groups, suppressed], ignore_index=True)
    return anonymized_data

anonymized_data = apply_k_anonymity(data, quasi_identifiers, k)
```

[6]:
```
# Step 5: Save the anonymized dataset
anonymized_data.to_csv("adult_dataset_anonymized.csv", index=False)

print("Anonymized dataset saved as 'adult_dataset_anonymized.csv'.")
```

```
Anonymized dataset saved as 'adult_dataset_anonymized.csv'.
```

[7]:
```
anonymized_data
```

[7]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per
0	3	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	
1	4	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	
2	2	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	
3	1	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	
4	2	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
32556	Suppressed	NaN	120478	Suppressed	11	Suppressed	NaN	Unmarried	Suppressed	Suppressed	0	0	
32557	Suppressed	Private	199655	Suppressed	14	Suppressed	Other-service	Not-in-family	Suppressed	Suppressed	0	0	
32558	Suppressed	Self-emp-not-inc	99359	Suppressed	15	Suppressed	Prof-specialty	Not-in-family	Suppressed	Suppressed	1086	0	

[7]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per
0	3	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	
1	4	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	
2	2	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	
3	1	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	
4	2	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
32556	Suppressed	NaN	120478	Suppressed	11	Suppressed	NaN	Unmarried	Suppressed	Suppressed	0	0	
32557	Suppressed	Private	199655	Suppressed	14	Suppressed	Other-service	Not-in-family	Suppressed	Suppressed	0	0	
32558	Suppressed	Self-emp-not-inc	99359	Suppressed	15	Suppressed	Prof-specialty	Not-in-family	Suppressed	Suppressed	1086	0	
32559	Suppressed	Private	34066	Suppressed	6	Suppressed	Handlers-cleaners	Husband	Suppressed	Suppressed	0	0	
32560	Suppressed	Private	116138	Suppressed	14	Suppressed	Tech-support	Not-in-family	Suppressed	Suppressed	0	0	

32561 rows × 15 columns

2. Implement (X,Y) -Linkability for the adult data set via code.

```
import pandas as pd

# Step 1: Load the dataset
columns = [
    "age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
    "occupation", "relationship", "race", "sex", "capital_gain",
    "capital_loss", "hours_per_week", "native_country", "income"
]

data = pd.read_csv("adult.data", names=columns, sep=r',\s*', engine='python', na_values='?')
```

```
data
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	

```
data
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-

32561 rows × 15 columns

[3]:
```
# Step 2: Define quasi-identifiers (X) and sensitive attributes (Y)
quasi_identifiers = ["age", "sex", "race"]
sensitive_attribute = "income"
```

[4]:
```
# Step 3: Group by quasi-identifiers and calculate diversity of sensitive attributes
def calculate_l_diversity(group, sensitive_column):
    # Calculate the number of unique values in the sensitive column
    return group[sensitive_column].nunique()

# Group data by quasi-identifiers
grouped = data.groupby(quasi_identifiers)

# Calculate diversity for each group
diversity = grouped[sensitive_attribute].nunique().reset_index(name='diversity')
```

[5]:
```
# Step 4: Ensure l-diversity (e.g., l = 2)
l = 2
sufficiently_diverse_groups = diversity[diversity['diversity'] >= l]

# Filter original data to keep only sufficiently diverse groups
diverse_data = data.merge(sufficiently_diverse_groups[quasi_identifiers], on=quasi_identifiers, how='inner')
```

[6]:
```
# Step 5: Save the (X, Y)-Linkable dataset
diverse_data.to_csv("adult_dataset_linkable.csv", index=False)

print("Linkable dataset saved as 'adult_dataset_linkable.csv'.")
```

```
Linkable dataset saved as 'adult_dataset_linkable.csv'.
```

[7]:
```
diverse_data
```

[7]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
29443	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
29444	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-
29445	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-
29446	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-
29447	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-

29448 rows × 15 columns