# Jaccad Coefficient (SMC),

## it's specifically useful for binary attributes where the focus is on the presence of attributes (ignoring the absence).

### Importing required Libraries

import pandas as pd  
import numpy as np  
from sklearn.preprocessing import LabelEncoder  
from sklearn.metrics import jaccard\_score

### Load Datasets

# Load datasets  
adult\_df = pd.read\_csv("../adult/adult\_trim.data", header=None) # No header  
titanic\_df = pd.read\_csv('../titanic/titanic\_trim.csv') # Has header  
  
# Rename columns for clarity  
adult\_df.columns = ["age", "workclass", "fnlwgt", "education", "education\_num",   
 "marital\_status", "occupation", "relationship", "race", "sex",   
 "capital\_gain", "capital\_loss", "hours\_per\_week", "native\_country", "income"]  
adult\_df.dropna(inplace=True)

adult\_df

age workclass fnlwgt education education\_num \  
0 39 State-gov 77516 Bachelors 13   
1 50 Self-emp-not-inc 83311 Bachelors 13   
2 38 Private 215646 HS-grad 9   
3 53 Private 234721 11th 7   
4 28 Private 338409 Bachelors 13   
.. ... ... ... ... ...   
95 29 Local-gov 115585 Some-college 10   
96 48 Self-emp-not-inc 191277 Doctorate 16   
97 37 Private 202683 Some-college 10   
98 48 Private 171095 Assoc-acdm 12   
99 32 Federal-gov 249409 HS-grad 9   
  
 marital\_status occupation relationship race sex \  
0 Never-married Adm-clerical Not-in-family White Male   
1 Married-civ-spouse Exec-managerial Husband White Male   
2 Divorced Handlers-cleaners Not-in-family White Male   
3 Married-civ-spouse Handlers-cleaners Husband Black Male   
4 Married-civ-spouse Prof-specialty Wife Black Female   
.. ... ... ... ... ...   
95 Never-married Handlers-cleaners Not-in-family White Male   
96 Married-civ-spouse Prof-specialty Husband White Male   
97 Married-civ-spouse Sales Husband White Male   
98 Divorced Exec-managerial Unmarried White Female   
99 Never-married Other-service Own-child Black Male   
  
 capital\_gain capital\_loss hours\_per\_week native\_country income   
0 2174 0 40 United-States <=50K   
1 0 0 13 United-States <=50K   
2 0 0 40 United-States <=50K   
3 0 0 40 United-States <=50K   
4 0 0 40 Cuba <=50K   
.. ... ... ... ... ...   
95 0 0 50 United-States <=50K   
96 0 1902 60 United-States >50K   
97 0 0 48 United-States >50K   
98 0 0 40 England <=50K   
99 0 0 40 United-States <=50K   
  
[100 rows x 15 columns]

titanic\_df

PassengerId Survived Pclass \  
0 1 0 3   
1 2 1 1   
2 3 1 3   
3 4 1 1   
4 5 0 3   
.. ... ... ...   
150 151 0 2   
151 152 1 1   
152 153 0 3   
153 154 0 3   
154 155 0 3   
  
 Name Sex Age SibSp \  
0 Braund, Mr. Owen Harris male 22.0 1   
1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1   
2 Heikkinen, Miss. Laina female 26.0 0   
3 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1   
4 Allen, Mr. William Henry male 35.0 0   
.. ... ... ... ...   
150 Bateman, Rev. Robert James male 51.0 0   
151 Pears, Mrs. Thomas (Edith Wearne) female 22.0 1   
152 Meo, Mr. Alfonzo male 55.5 0   
153 van Billiard, Mr. Austin Blyler male 40.5 0   
154 Olsen, Mr. Ole Martin male NaN 0   
  
 Parch Ticket Fare Cabin Embarked   
0 0 A/5 21171 7.2500 NaN S   
1 0 PC 17599 71.2833 C85 C   
2 0 STON/O2. 3101282 7.9250 NaN S   
3 0 113803 53.1000 C123 S   
4 0 373450 8.0500 NaN S   
.. ... ... ... ... ...   
150 0 S.O.P. 1166 12.5250 NaN S   
151 0 113776 66.6000 C2 S   
152 0 A.5. 11206 8.0500 NaN S   
153 2 A/5. 851 14.5000 NaN S   
154 0 Fa 265302 7.3125 NaN S   
  
[155 rows x 12 columns]

### Select relevant columns from Adult dataset (mix of nominal and ratio-scaled)

adult\_df = adult\_df[["age", "workclass", "education", "education\_num", "sex"]]  
  
adult\_df

age workclass education education\_num sex  
0 39 State-gov Bachelors 13 Male  
1 50 Self-emp-not-inc Bachelors 13 Male  
2 38 Private HS-grad 9 Male  
3 53 Private 11th 7 Male  
4 28 Private Bachelors 13 Female  
.. ... ... ... ... ...  
95 29 Local-gov Some-college 10 Male  
96 48 Self-emp-not-inc Doctorate 16 Male  
97 37 Private Some-college 10 Male  
98 48 Private Assoc-acdm 12 Female  
99 32 Federal-gov HS-grad 9 Male  
  
[100 rows x 5 columns]

### Encode nominal attributes as integers for processing

label\_encoders = {}  
for column in adult\_df.columns:  
 if adult\_df[column].dtype == object:  
 le = LabelEncoder()  
 adult\_df[column] = le.fit\_transform(adult\_df[column])  
 label\_encoders[column] = le  
  
adult\_df

### Clean and preprocess Titanic dataset

titanic\_df.dropna(inplace=True)  
titanic\_df

PassengerId Survived Pclass \  
1 2 1 1   
3 4 1 1   
6 7 0 1   
10 11 1 3   
11 12 1 1   
21 22 1 2   
23 24 1 1   
27 28 0 1   
52 53 1 1   
54 55 0 1   
62 63 0 1   
66 67 1 2   
75 76 0 3   
88 89 1 1   
92 93 0 1   
96 97 0 1   
97 98 1 1   
102 103 0 1   
110 111 0 1   
118 119 0 1   
123 124 1 2   
124 125 0 1   
136 137 1 1   
137 138 0 1   
139 140 0 1   
148 149 0 2   
151 152 1 1   
  
 Name Sex Age SibSp \  
1 Cumings, Mrs. John Bradley (Florence Briggs Th... female 38.0 1   
3 Futrelle, Mrs. Jacques Heath (Lily May Peel) female 35.0 1   
6 McCarthy, Mr. Timothy J male 54.0 0   
10 Sandstrom, Miss. Marguerite Rut female 4.0 1   
11 Bonnell, Miss. Elizabeth female 58.0 0   
21 Beesley, Mr. Lawrence male 34.0 0   
23 Sloper, Mr. William Thompson male 28.0 0   
27 Fortune, Mr. Charles Alexander male 19.0 3   
52 Harper, Mrs. Henry Sleeper (Myna Haxtun) female 49.0 1   
54 Ostby, Mr. Engelhart Cornelius male 65.0 0   
62 Harris, Mr. Henry Birkhardt male 45.0 1   
66 Nye, Mrs. (Elizabeth Ramell) female 29.0 0   
75 Moen, Mr. Sigurd Hansen male 25.0 0   
88 Fortune, Miss. Mabel Helen female 23.0 3   
92 Chaffee, Mr. Herbert Fuller male 46.0 1   
96 Goldschmidt, Mr. George B male 71.0 0   
97 Greenfield, Mr. William Bertram male 23.0 0   
102 White, Mr. Richard Frasar male 21.0 0   
110 Porter, Mr. Walter Chamberlain male 47.0 0   
118 Baxter, Mr. Quigg Edmond male 24.0 0   
123 Webber, Miss. Susan female 32.5 0   
124 White, Mr. Percival Wayland male 54.0 0   
136 Newsom, Miss. Helen Monypeny female 19.0 0   
137 Futrelle, Mr. Jacques Heath male 37.0 1   
139 Giglio, Mr. Victor male 24.0 0   
148 Navratil, Mr. Michel ("Louis M Hoffman") male 36.5 0   
151 Pears, Mrs. Thomas (Edith Wearne) female 22.0 1   
  
 Parch Ticket Fare Cabin Embarked   
1 0 PC 17599 71.2833 C85 C   
3 0 113803 53.1000 C123 S   
6 0 17463 51.8625 E46 S   
10 1 PP 9549 16.7000 G6 S   
11 0 113783 26.5500 C103 S   
21 0 248698 13.0000 D56 S   
23 0 113788 35.5000 A6 S   
27 2 19950 263.0000 C23 C25 C27 S   
52 0 PC 17572 76.7292 D33 C   
54 1 113509 61.9792 B30 C   
62 0 36973 83.4750 C83 S   
66 0 C.A. 29395 10.5000 F33 S   
75 0 348123 7.6500 F G73 S   
88 2 19950 263.0000 C23 C25 C27 S   
92 0 W.E.P. 5734 61.1750 E31 S   
96 0 PC 17754 34.6542 A5 C   
97 1 PC 17759 63.3583 D10 D12 C   
102 1 35281 77.2875 D26 S   
110 0 110465 52.0000 C110 S   
118 1 PC 17558 247.5208 B58 B60 C   
123 0 27267 13.0000 E101 S   
124 1 35281 77.2875 D26 S   
136 2 11752 26.2833 D47 S   
137 0 113803 53.1000 C123 S   
139 0 PC 17593 79.2000 B86 C   
148 2 230080 26.0000 F2 S   
151 0 113776 66.6000 C2 S

### Select relevant columns from Titanic dataset (mix of nominal and ratio-scaled)

titanic\_df = titanic\_df[["Age", "Sex", "Pclass", "Fare", "Embarked"]]  
titanic\_df

Age Sex Pclass Fare Embarked  
1 38.0 female 1 71.2833 C  
3 35.0 female 1 53.1000 S  
6 54.0 male 1 51.8625 S  
10 4.0 female 3 16.7000 S  
11 58.0 female 1 26.5500 S  
21 34.0 male 2 13.0000 S  
23 28.0 male 1 35.5000 S  
27 19.0 male 1 263.0000 S  
52 49.0 female 1 76.7292 C  
54 65.0 male 1 61.9792 C  
62 45.0 male 1 83.4750 S  
66 29.0 female 2 10.5000 S  
75 25.0 male 3 7.6500 S  
88 23.0 female 1 263.0000 S  
92 46.0 male 1 61.1750 S  
96 71.0 male 1 34.6542 C  
97 23.0 male 1 63.3583 C  
102 21.0 male 1 77.2875 S  
110 47.0 male 1 52.0000 S  
118 24.0 male 1 247.5208 C  
123 32.5 female 2 13.0000 S  
124 54.0 male 1 77.2875 S  
136 19.0 female 1 26.2833 S  
137 37.0 male 1 53.1000 S  
139 24.0 male 1 79.2000 C  
148 36.5 male 2 26.0000 S  
151 22.0 female 1 66.6000 S

### Encode Nominal as Integers for processing

label\_encoders\_titanic = {}  
for column in titanic\_df.columns:  
 if titanic\_df[column].dtype == object:  
 le = LabelEncoder()  
 titanic\_df[column] = le.fit\_transform(titanic\_df[column])  
 label\_encoders[column] = le  
  
titanic\_df

C:\Users\debat\AppData\Local\Temp\ipykernel\_16092\3305425594.py:5: SettingWithCopyWarning:   
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead  
  
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy  
 titanic\_df[column] = le.fit\_transform(titanic\_df[column])  
C:\Users\debat\AppData\Local\Temp\ipykernel\_16092\3305425594.py:5: SettingWithCopyWarning:   
A value is trying to be set on a copy of a slice from a DataFrame.  
Try using .loc[row\_indexer,col\_indexer] = value instead  
  
See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user\_guide/indexing.html#returning-a-view-versus-a-copy  
 titanic\_df[column] = le.fit\_transform(titanic\_df[column])

Age Sex Pclass Fare Embarked  
1 38.0 0 1 71.2833 0  
3 35.0 0 1 53.1000 1  
6 54.0 1 1 51.8625 1  
10 4.0 0 3 16.7000 1  
11 58.0 0 1 26.5500 1  
21 34.0 1 2 13.0000 1  
23 28.0 1 1 35.5000 1  
27 19.0 1 1 263.0000 1  
52 49.0 0 1 76.7292 0  
54 65.0 1 1 61.9792 0  
62 45.0 1 1 83.4750 1  
66 29.0 0 2 10.5000 1  
75 25.0 1 3 7.6500 1  
88 23.0 0 1 263.0000 1  
92 46.0 1 1 61.1750 1  
96 71.0 1 1 34.6542 0  
97 23.0 1 1 63.3583 0  
102 21.0 1 1 77.2875 1  
110 47.0 1 1 52.0000 1  
118 24.0 1 1 247.5208 0  
123 32.5 0 2 13.0000 1  
124 54.0 1 1 77.2875 1  
136 19.0 0 1 26.2833 1  
137 37.0 1 1 53.1000 1  
139 24.0 1 1 79.2000 0  
148 36.5 1 2 26.0000 1  
151 22.0 0 1 66.6000 1

### Combine the datasets into a list for further processing

# Combine the datasets into a list for further processing  
datasets = {  
 "Adult Dataset": adult\_df,  
 "Titanic Dataset": titanic\_df  
}

### Computing Jaccard Coefficient

#### The Jaccard Coefficient is calculated as the size of the intersection divided by the size of the union of the attribute sets. It's typically used for binary or nominal data.

def jaccard\_coefficient(a, b):  
 """Calculate the Jaccard Coefficient between two vectors."""  
 try:  
 # Convert to binary format if necessary (for categorical data)  
 return jaccard\_score(a, b, average='macro')  
 except Exception as e:  
 return np.nan  
  
# Function to create the Jaccard proximity matrix  
def calculate\_jaccard\_matrix(dataset):  
 n = len(dataset)  
 jaccard\_matrix = np.zeros((n, n))  
   
 for i in range(n):  
 # print(f"{i}/{n}")  
 for j in range(n):  
 jaccard\_matrix[i, j] = jaccard\_coefficient(dataset.iloc[i].values, dataset.iloc[j].values)  
   
 return pd.DataFrame(jaccard\_matrix)

### Calculating Matrices for the data sets

jaccard\_matrix\_adult = calculate\_jaccard\_matrix(adult\_df)  
jaccard\_matrix\_titanic = calculate\_jaccard\_matrix(titanic\_df)

### Displaying the Matrices

#### Adult Dataset: Jaccard Matrix

jaccard\_matrix\_adult

0 1 2 3 4 5 6 \  
0 1.000000 0.428571 0.125000 0.071429 0.250000 0.000000 0.000000   
1 0.428571 1.000000 0.125000 0.071429 0.250000 0.000000 0.000000   
2 0.125000 0.125000 1.000000 0.250000 0.125000 0.125000 0.125000   
3 0.071429 0.071429 0.250000 1.000000 0.142857 0.125000 0.125000   
4 0.250000 0.250000 0.125000 0.142857 1.000000 0.250000 0.250000   
.. ... ... ... ... ... ... ...   
95 0.111111 0.111111 0.125000 0.062500 0.000000 0.000000 0.000000   
96 0.111111 0.250000 0.125000 0.062500 0.000000 0.000000 0.000000   
97 0.111111 0.111111 0.285714 0.214286 0.111111 0.285714 0.111111   
98 0.000000 0.000000 0.125000 0.125000 0.250000 0.250000 0.285714   
99 0.071429 0.071429 0.300000 0.055556 0.000000 0.000000 0.000000   
  
 7 8 9 ... 90 91 92 93 \  
0 0.125000 0.000 0.428571 ... 0.125000 0.000000 0.000000 0.000000   
1 0.285714 0.000 0.428571 ... 0.111111 0.000000 0.000000 0.000000   
2 0.333333 0.125 0.285714 ... 0.285714 0.125000 0.125000 0.333333   
3 0.071429 0.125 0.250000 ... 0.214286 0.125000 0.125000 0.142857   
4 0.000000 0.250 0.428571 ... 0.111111 0.250000 0.428571 0.285714   
.. ... ... ... ... ... ... ... ...   
95 0.125000 0.000 0.111111 ... 0.111111 0.250000 0.250000 0.000000   
96 0.285714 0.000 0.111111 ... 0.111111 0.000000 0.000000 0.000000   
97 0.125000 0.125 0.250000 ... 0.250000 0.666667 0.428571 0.125000   
98 0.000000 0.250 0.111111 ... 0.111111 0.285714 0.285714 0.285714   
99 0.300000 0.000 0.071429 ... 0.071429 0.000000 0.000000 0.166667   
  
 94 95 96 97 98 99   
0 0.428571 0.111111 0.111111 0.111111 0.000 0.071429   
1 0.428571 0.111111 0.250000 0.111111 0.000 0.071429   
2 0.125000 0.125000 0.125000 0.285714 0.125 0.300000   
3 0.071429 0.062500 0.062500 0.214286 0.125 0.055556   
4 0.250000 0.000000 0.000000 0.111111 0.250 0.000000   
.. ... ... ... ... ... ...   
95 0.250000 1.000000 0.111111 0.428571 0.000 0.071429   
96 0.111111 0.111111 1.000000 0.111111 0.125 0.071429   
97 0.111111 0.428571 0.111111 1.000000 0.125 0.071429   
98 0.000000 0.000000 0.125000 0.125000 1.000 0.000000   
99 0.071429 0.071429 0.071429 0.071429 0.000 1.000000   
  
[100 rows x 100 columns]

#### Titanic Dataset: Jaccard Matrix

jaccard\_matrix\_titanic

0 1 2 3 4 5 6 7 8 9 ... 17 18 19 \  
0 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
1 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
2 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
3 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
4 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
5 NaN NaN NaN NaN NaN 1.000000 NaN 0.111111 NaN NaN ... NaN 0.111111 NaN   
6 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
7 NaN NaN NaN NaN NaN 0.111111 NaN 1.000000 NaN NaN ... NaN 0.200000 NaN   
8 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
9 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
10 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
11 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
12 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
13 NaN NaN NaN NaN NaN 0.047619 NaN 0.333333 NaN NaN ... NaN 0.111111 NaN   
14 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
15 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
16 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
17 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
18 NaN NaN NaN NaN NaN 0.111111 NaN 0.200000 NaN NaN ... NaN 1.000000 NaN   
19 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
20 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
21 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
22 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
23 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
24 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
25 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
26 NaN NaN NaN NaN NaN NaN NaN NaN NaN NaN ... NaN NaN NaN   
  
 20 21 22 23 24 25 26   
0 NaN NaN NaN NaN NaN NaN NaN   
1 NaN NaN NaN NaN NaN NaN NaN   
2 NaN NaN NaN NaN NaN NaN NaN   
3 NaN NaN NaN NaN NaN NaN NaN   
4 NaN NaN NaN NaN NaN NaN NaN   
5 NaN NaN NaN NaN NaN NaN NaN   
6 NaN NaN NaN NaN NaN NaN NaN   
7 NaN NaN NaN NaN NaN NaN NaN   
8 NaN NaN NaN NaN NaN NaN NaN   
9 NaN NaN NaN NaN NaN NaN NaN   
10 NaN NaN NaN NaN NaN NaN NaN   
11 NaN NaN NaN NaN NaN NaN NaN   
12 NaN NaN NaN NaN NaN NaN NaN   
13 NaN NaN NaN NaN NaN NaN NaN   
14 NaN NaN NaN NaN NaN NaN NaN   
15 NaN NaN NaN NaN NaN NaN NaN   
16 NaN NaN NaN NaN NaN NaN NaN   
17 NaN NaN NaN NaN NaN NaN NaN   
18 NaN NaN NaN NaN NaN NaN NaN   
19 NaN NaN NaN NaN NaN NaN NaN   
20 NaN NaN NaN NaN NaN NaN NaN   
21 NaN NaN NaN NaN NaN NaN NaN   
22 NaN NaN NaN NaN NaN NaN NaN   
23 NaN NaN NaN NaN NaN NaN NaN   
24 NaN NaN NaN NaN NaN NaN NaN   
25 NaN NaN NaN NaN NaN NaN NaN   
26 NaN NaN NaN NaN NaN NaN NaN   
  
[27 rows x 27 columns]

### Explanation

Jaccard Coefficient Calculation: This metric focuses on the presence of attributes, making it useful for binary data. It calculates the ratio of the intersection of attributes to the union of attributes between two data points.

Handling Nominal Data: The Jaccard Coefficient is generally used for binary data, but in this case, we can apply it to the encoded nominal attributes as they are binary representations of categories.

### Observation and Analysis

The resulting matrices will have values between 0 and 1, where 1 indicates identical presence/absence of attributes (perfect similarity) and 0 indicates no similarity.

Jaccard Coefficient is particularly sensitive to the presence of attributes (1s) and ignores cases where both attributes are absent (0s), making it a useful measure for sparse datasets.