Data Processing and Cleaning

```
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.linear_model import LogisticRegression

konsumsi_rokok=pd.read_csv(r"C:\Users\muham\OneDrive\Dokumen\DATA
ANALYST\database\susenas\logistik python\fix.csv")
konsumsi_rokok
```

```
      rokok  pend  usia  kapita kelamin
0         1     1     1       1    male
1         1     1     1       0    male
2         0     1     0       0    male
3         0     1     0       0    male
4         1     1     1       0    male
...     ...   ...   ...     ...     ...
2399      1     1     0       0    male
2400      1     1     0       0    male
2401      1     1     0       0    male
2402      0     1     0       0    male
2403      1     1     0       0    male

[2404 rows x 5 columns]
```

```
len(konsumsi_rokok)
```

```
2404
```

```
konsumsi_rokok.head(10)
```

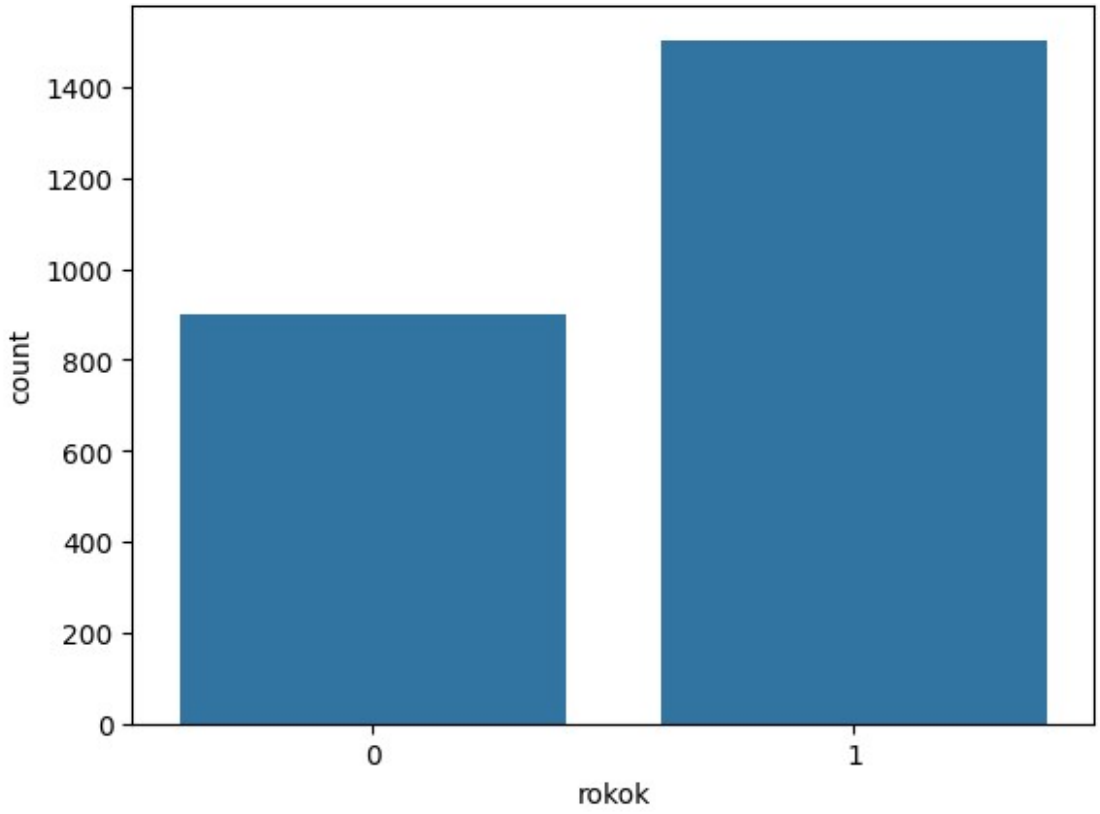
```
   rokok  pend  usia  kapita kelamin
0      1     1     1       1    male
1      1     1     1       0    male
2      0     1     0       0    male
3      0     1     0       0    male
4      1     1     1       0    male
5      1     1     1       1    male
6      1     1     0       0    male
7      1     1     1       0    male
8      0     1     0       0    male
9      0     1     0       0    male
```

```
konsumsi_rokok.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2404 entries, 0 to 2403
Data columns (total 5 columns):
 #   Column   Non-Null Count  Dtype 
---  ------   --------------  ----- 
 0   rokok    2404 non-null   int64 
 1   pend     2404 non-null   int64 
 2   usia     2404 non-null   int64 
 3   kapita   2404 non-null   int64 
 4   kelamin  2404 non-null   object
dtypes: int64(4), object(1)
memory usage: 94.0+ KB
```

```
sns.countplot(x='rokok',data=konsumsi_rokok)
```

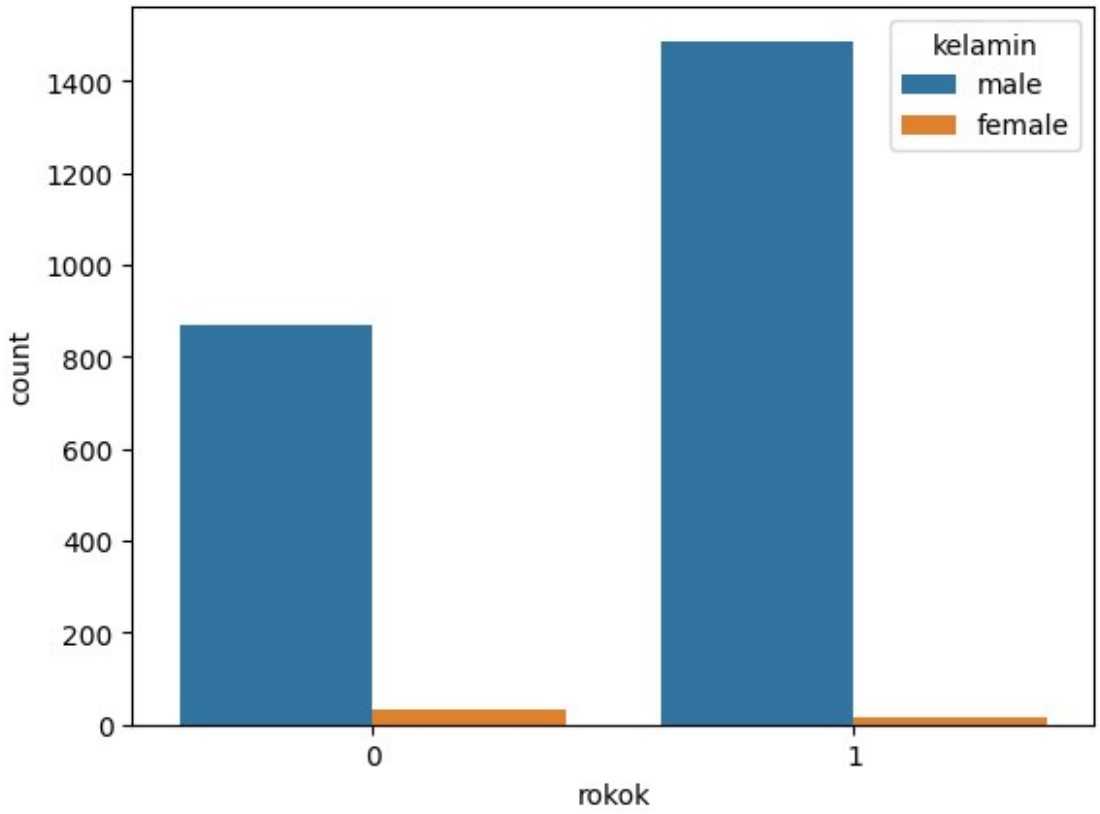
```
<Axes: xlabel='rokok', ylabel='count'>
```



pembagian data untuk perokok yaitu 0 adalah untuk perokok ringan dan sedang yaitu <=104 batang perbulannya. Sedangkan untuk data 1 adalah perokok berat yang merokok >=105 batang rokok perbulannya.

```
sns.countplot(x='rokok',data=konsumsi_rokok, hue= 'kelamin')
```

```
<Axes: xlabel='rokok', ylabel='count'>
```



Removing Null Data From Our Data Set Ini akan menghasilkan DataFrame nilai boolean yang berisi sel Truejika nilainya nol dan Falsesebaliknya. Berikut adalah gambar tampilannya:

```
konsumsi_rokok.isna()
```

```
      rokok   pend   usia  kapita  kelamin
0     False  False  False   False    False
1     False  False  False   False    False
2     False  False  False   False    False
3     False  False  False   False    False
4     False  False  False   False    False
...     ...    ...    ...     ...      ...
2399  False  False  False   False    False
2400  False  False  False   False    False
2401  False  False  False   False    False
2402  False  False  False   False    False
2403  False  False  False   False    False

[2404 rows x 5 columns]
```

Adding Dummy Variables to the pandas Data Frame

```
konsumsi_rokok['sex'] = konsumsi_rokok['kelamin'].apply(lambda x: 1 if
x== 'male' else 0)
```

```
konsumsi_rokok.describe()
```

	rokok	pend	usia	kapita	sex
count	2404.000000	2404.000000	2404.000000	2404.000000	2404.000000
mean	0.625624	0.950499	0.374376	0.025374	0.979201
std	0.484062	0.216956	0.484062	0.157292	0.142739
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	1.000000	0.000000	0.000000	1.000000
50%	1.000000	1.000000	0.000000	0.000000	1.000000
75%	1.000000	1.000000	1.000000	0.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000

cleaning data

```
del(konsumsi_rokok['kelamin'])

konsumsi_rokok
```

	rokok	pend	usia	kapita	sex
0	1	1	1	1	1
1	1	1	1	0	1
2	0	1	0	0	1
3	0	1	0	0	1
4	1	1	1	0	1
...	...	...	...	...	...
2399	1	1	0	0	1
2400	1	1	0	0	1
2401	1	1	0	0	1
2402	0	1	0	0	1
2403	1	1	0	0	1

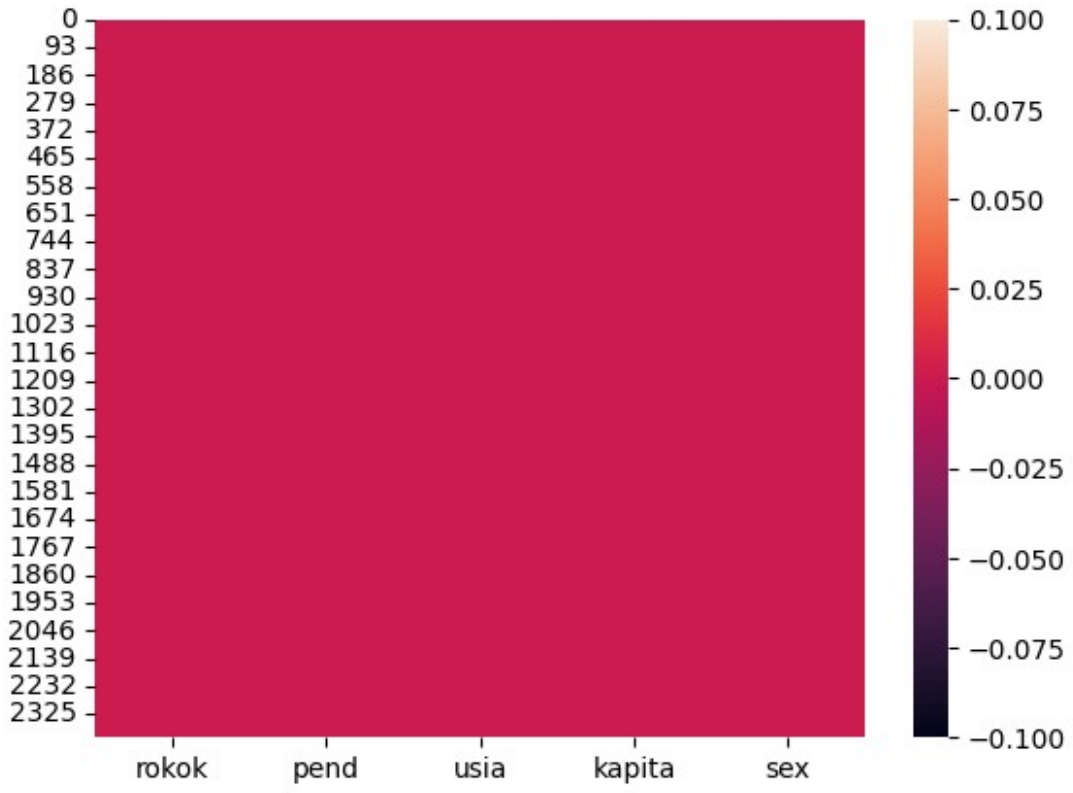
```
[2404 rows x 5 columns]
```

Definisi Operasional Variabel:

1. rokok : Konsumsi Rokok 1 = Merokok berat (>=105batang/bulan) 0 = Merokok sedang <=104 batang/bulan)
2. pend : Tingkat Pendidikan 1 = SD, SMP, SMA (sederajat) 0 = Diploma dan sarjana seterusnya
3. usia : Umur 1 = 15 - 35 tahun 0 = >= 36 tahun
4. kapita : Pendapatan 1 = Dibawah Rp.440.538,- / bulan (GK Sumatera Barat) 0 = diatas Rp.440.538,-
5. sex : Jenis Kelamin 1 = Laki-laki 0 = Perempuan

```
sns.heatmap(konsumsi_rokok.isna())
```

```
<Axes: >
```



Modelling

y adalah variabel Independen x adalah variabel Dependen

```
x=konsumsi_rokok[['sex', 'pend', 'usia', 'kapita']]
y=konsumsi_rokok['rokok']

y
```

```
0       1
1       1
2       0
3       0
4       1
       ..
2399    1
2400    1
2401    1
2402    0
2403    1
Name: rokok, Length: 2404, dtype: int64
```

```
x
```

```
      sex  pend  usia  kapita
0       1     1     1       1
1       1     1     1       0
2       1     1     0       0
3       1     1     0       0
4       1     1     1       0
...   ...   ...   ...     ...
2399    1     1     0       0
2400    1     1     0       0
2401    1     1     0       0
2402    1     1     0       0
2403    1     1     0       0

[2404 rows x 4 columns]
```

Training the Logistic Regression Model

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.33, random_state=42)

lr=LogisticRegression()

lr.fit(x_train,y_train)

LogisticRegression()
```

Testing

See how our model is performing

```
y_pred = lr.predict (x_test)

result = pd.DataFrame({'Actual' : y_test, 'Predicted' : y_pred})
result
```

```
      Actual  Predicted
296        0          1
2239       0          1
787        1          1
211        0          1
532        1          1
...      ...        ...
393        1          1
1287       0          1
534        0          1
```

```
30          1           1
1176        1           1

[794 rows x 2 columns]
```

```
lr.coef_
```

```
array([[ 0.99964882, -0.1238021 ,  0.05457995, -0.10663223]])
```

```
lr.intercept_
```

```
array([-0.36611589])
```

Interpretasi hasil:

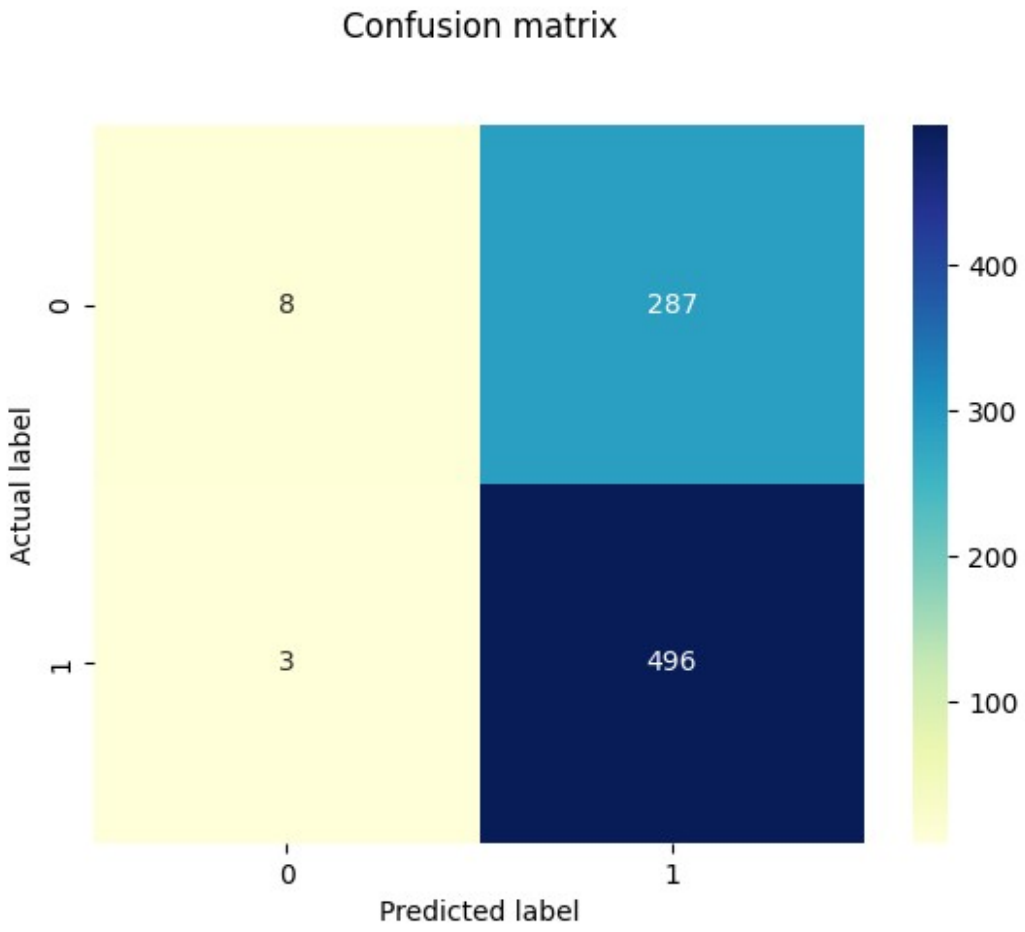
1. Pada variabel jenis kelamin untuk setiap peningkatan satu unit dalam rata" peringkat atau (dalam satu kesatuan) akan menyebabkan peningkatan konsumsi rokok oleh pria sebesar 0.99964882.
2. Pada Variabel Pendidikan Menunjukan bahwa tingkat pendidikan individu memiliki dampak negatif probabilitas seseorang menjadi perokok berat -0.1238021.
3. pada variabel umur Hal ini diinterpretasikan bahwa individu yang memiliki umur rentang 15-35 tahun memiliki peluang merokok lebih besar 0.05457995 dibanding individu yang memiliki umur besar sama dari 36 tahun.
4. Pada Variabel pendapatan pada garis kemiskinan kemungkinan untuk merokoknya lebih rendah -0.10663223.

```
from sklearn.metrics import confusion_matrix

cf_matrix = confusion_matrix(y_test, y_pred)

cf_matrix
```

```
array([[  8, 287],
       [  3, 496]], dtype=int64)
```

```
sns.heatmap(pd.DataFrame(cf_matrix), annot=True, cmap="YlGnBu",
fmt='g')
plt.title ('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

```
Text(0.5, 23.52222222222222, 'Predicted label')
```




```
from sklearn.metrics import classification_report
target_names = ['perokok ringan', 'perokok berat']
print(classification_report(y_test, y_pred,
target_names=target_names))
```

```
                precision    recall  f1-score   support

perokok ringan       0.73      0.03      0.05       295
 perokok berat       0.63      0.99      0.77       499

      accuracy                           0.63       794
     macro avg       0.68      0.51      0.41       794
  weighted avg       0.67      0.63      0.51       794
```