Analisis Data

Dilakukan analisis data dengan memanfaatkan histogram untuk melihat apakah suatu atribut dapat digunakan sebagai pembeda. Atribut yang dinilai dapat digunakan sebagai pembeda adalah atribut yang dominan di nilai tertentu

Load & Preprocess Data

In [18]:

```
# Library Import
import pandas as pd
from pandas import DataFrame
import graphviz
from sklearn import preprocessing
import pickle

# Algorithm
from sklearn.naive_bayes import GaussianNB
from sklearn import tree
from sklearn.neighbors import KNeighborsClassifier
from sklearn.neural_network import MLPClassifier

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import KFold
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import confusion_matrix

from collections import defaultdict
import matplotlib.pyplot as plt
```

In [19]:

```
atributeNames = ["age","workclass","fnlwgt","education","education-num","marital-status","occupation","relationship","race","sex","capital-gain","capital-loss","hours-per-week","native-country","target"]
income = pd.read_csv('data/CensusIncome.data.csv', header=None, names = atributeNames, sep = ",\s", engine="python", na_values="?");
```

In [20]:

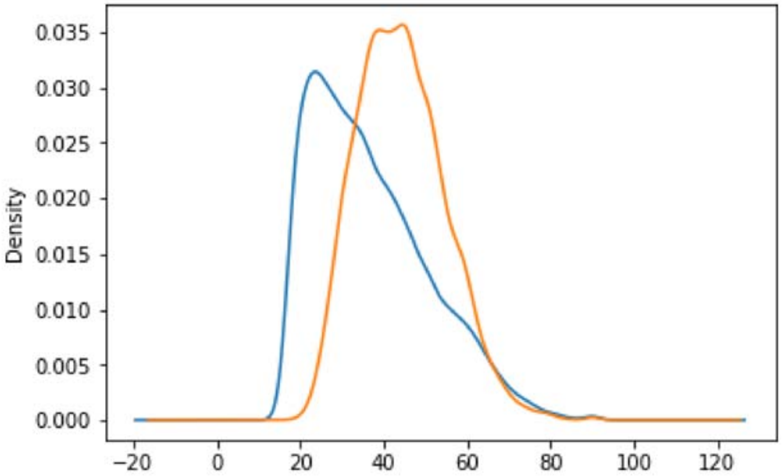
```
label = defaultdict(LabelEncoder)
income = income.fillna("NaN")
income = income.apply(lambda x: x if x.dtype != 'O' else label[x.name].fit_transform(x))
```

Histogram Analysis

Age Histogram

```
In [21]: income.groupby("target").age.plot(kind="kde")

         plt.figure();
         plt.show()
```

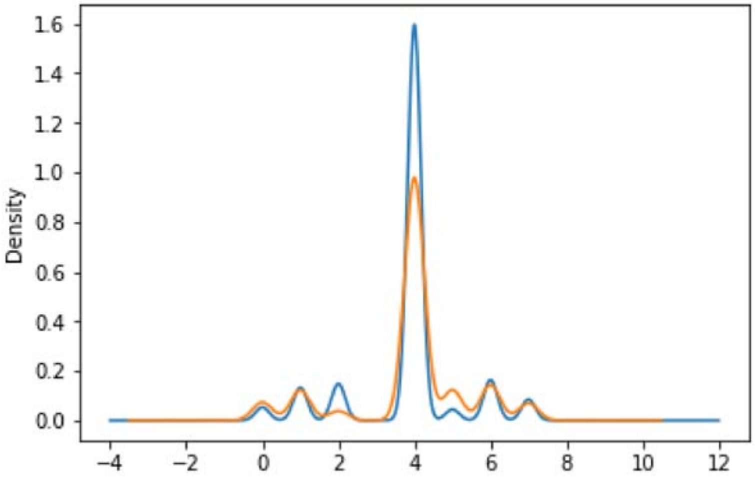


```
<matplotlib.figure.Figure at 0x7f656264db38>
```

Workclass Histogram

```
In [22]: income.groupby("target").workclass.plot(kind="kde")

         plt.figure()
         plt.show()
```

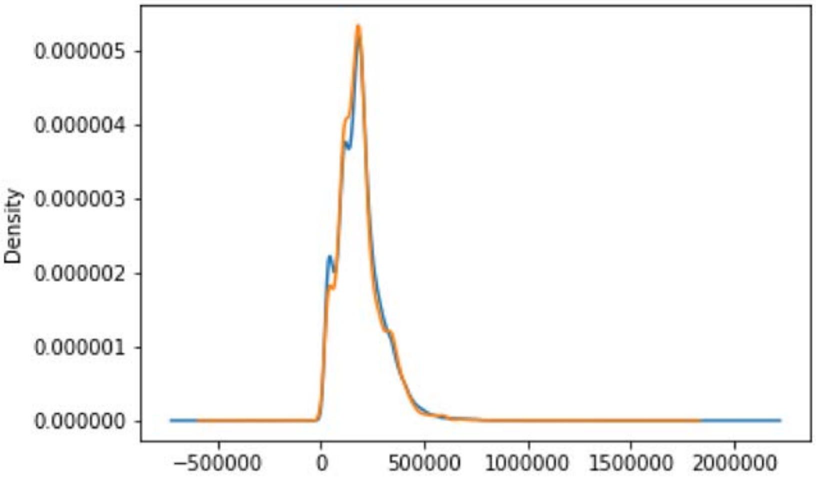


```
<matplotlib.figure.Figure at 0x7f65624889b0>
```

Final Weight Histogram

```
In [23]: income.groupby("target").fnlwgt.plot(kind="kde")

         plt.figure()
         plt.show()
```

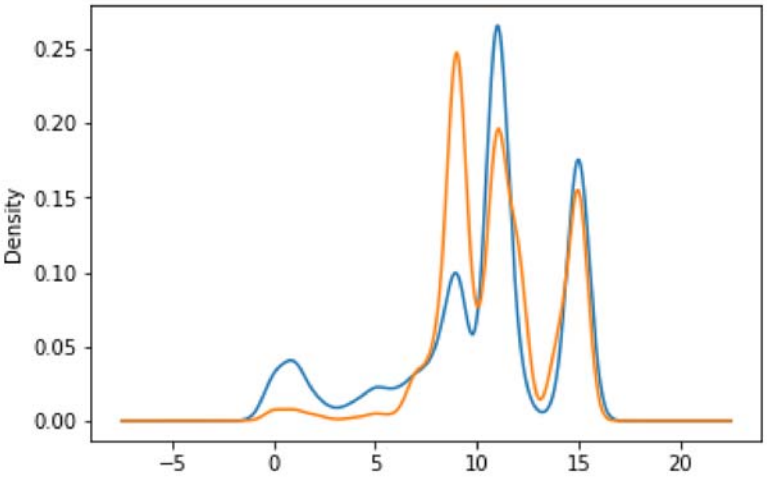


```
<matplotlib.figure.Figure at 0x7f6562503f98>
```

Education Histogram

```
In [24]: income.groupby("target").education.plot(kind="kde")

         plt.figure()
         plt.show()
```

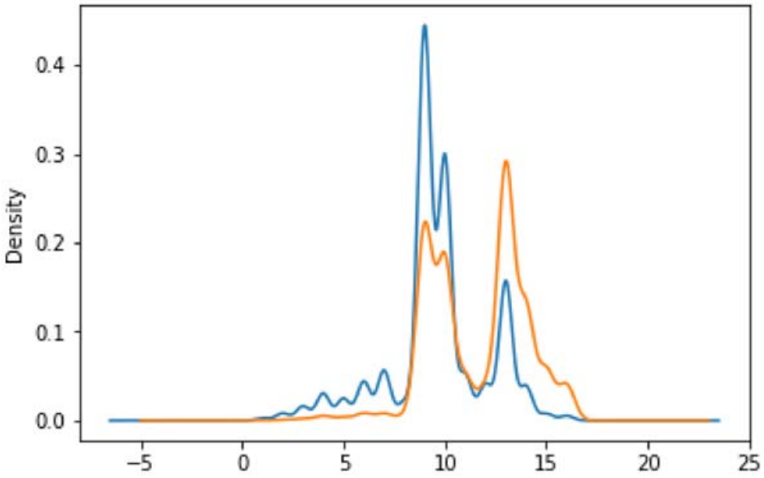


```
<matplotlib.figure.Figure at 0x7f6562006940>
```

Education Num Histogram

```
In [25]: income.groupby("target")["education-num"].plot(kind="kde")

         plt.figure()
         plt.show()
```

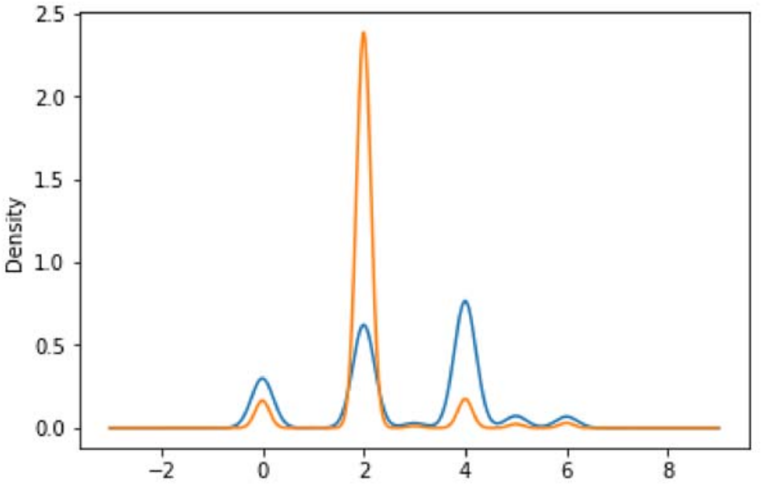


```
<matplotlib.figure.Figure at 0x7f65622aa4e0>
```

Marital Status Histogram

```
In [26]: income.groupby("target")["marital-status"].plot(kind="kde")

         plt.figure()
         plt.show()
```

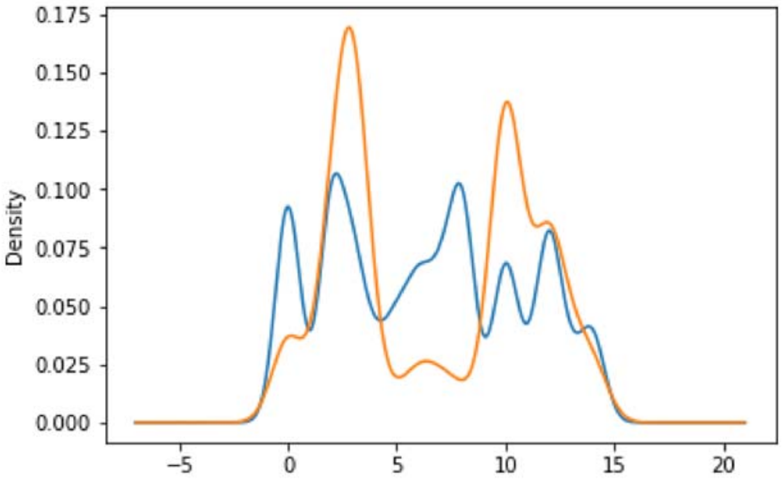


```
<matplotlib.figure.Figure at 0x7f65624fa668>
```

Occupation Histogram

```
In [27]: income.groupby("target")["occupation"].plot(kind="kde")

         plt.figure()
         plt.show()
```

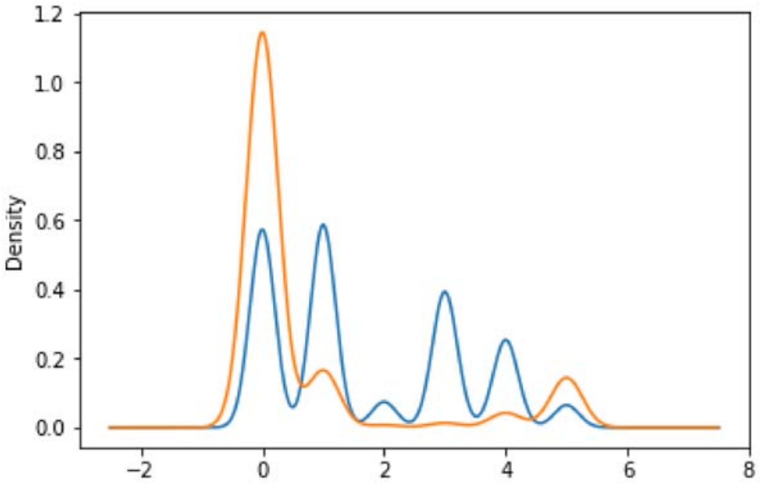


```
<matplotlib.figure.Figure at 0x7f65622aaa20>
```

Relationship Histogram

```
In [28]: income.groupby("target")["relationship"].plot(kind="kde")

         plt.figure()
         plt.show()
```

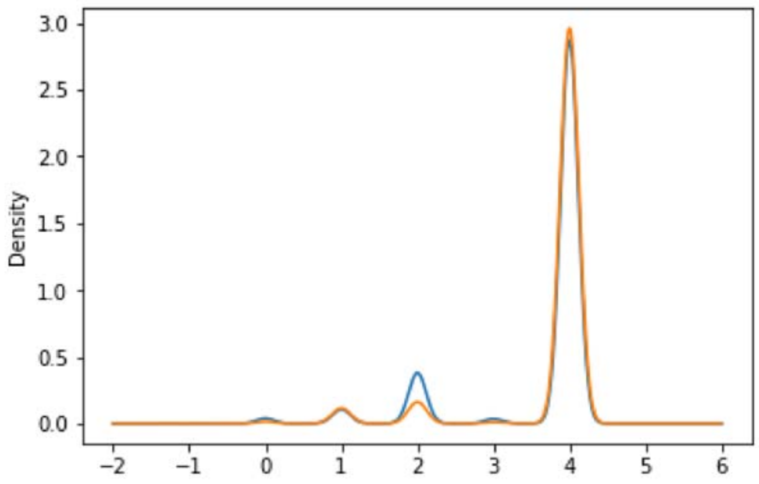


```
<matplotlib.figure.Figure at 0x7f65a46e9d68>
```

Race Histogram

In [29]:
```
income.groupby("target")["race"].plot(kind="kde")

plt.figure()
plt.show()
```

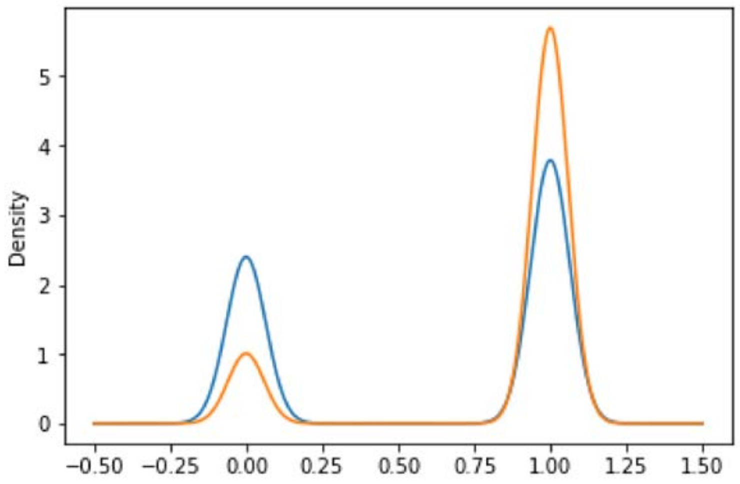


```
<matplotlib.figure.Figure at 0x7f656251c0f0>
```

Sex Histogram

In [30]:
```
income.groupby("target")["sex"].plot(kind="kde")

plt.figure()
plt.show()
```



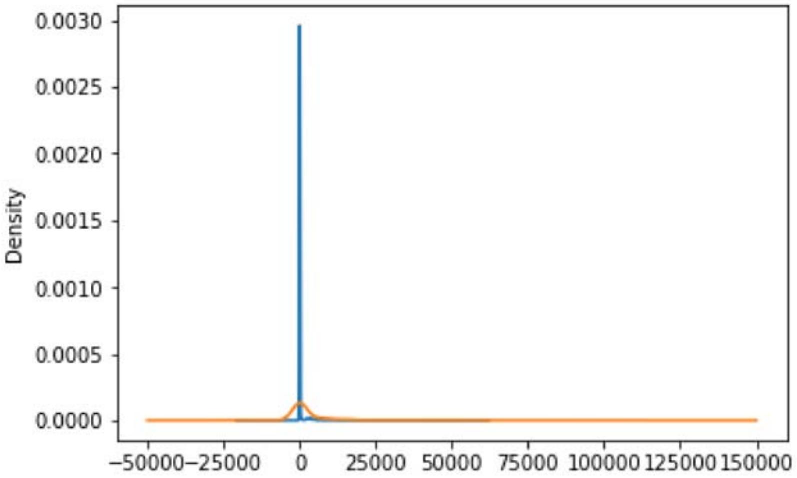
```
<matplotlib.figure.Figure at 0x7f6561c61ba8>
```

Capital Gain Histogram

In [31]:

```
income.groupby("target")["capital-gain"].plot(kind="kde")

plt.figure()
plt.show()
```



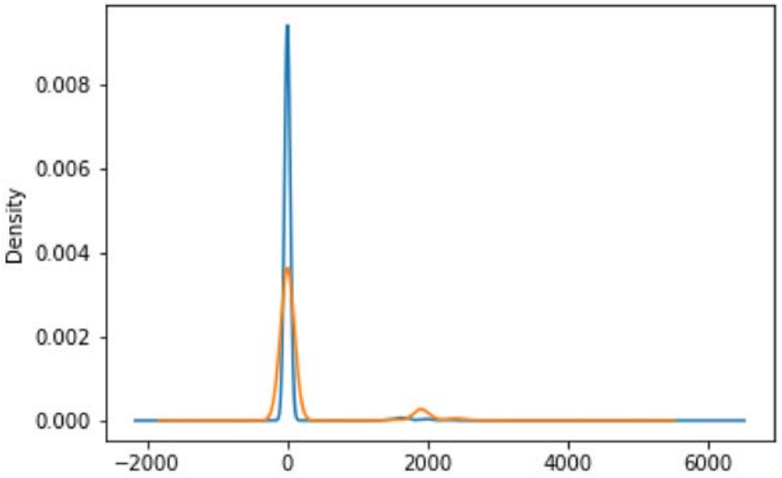
```
<matplotlib.figure.Figure at 0x7f656264d1d0>
```

Capital Loss Histogram

In [32]:

```
income.groupby("target")["capital-loss"].plot(kind="kde")

plt.figure()
plt.show()
```



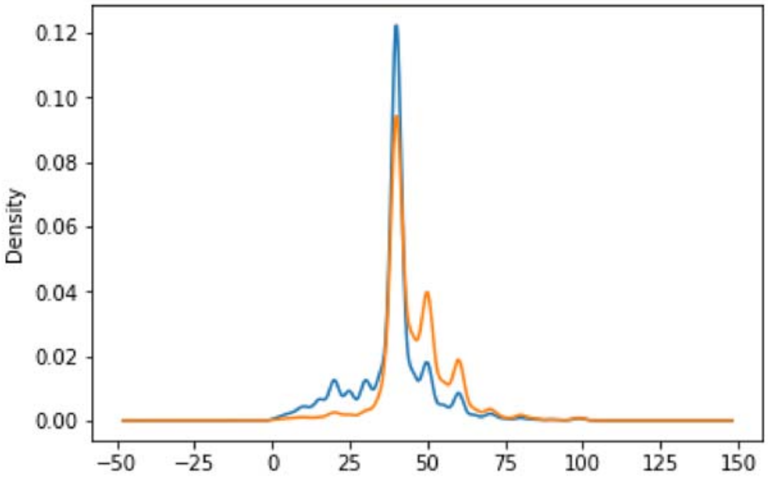
```
<matplotlib.figure.Figure at 0x7f656248e7b8>
```

Hours per Week Histogram

In [33]:

```
income.groupby("target")["hours-per-week"].plot(kind="kde")

plt.figure()
plt.show()
```



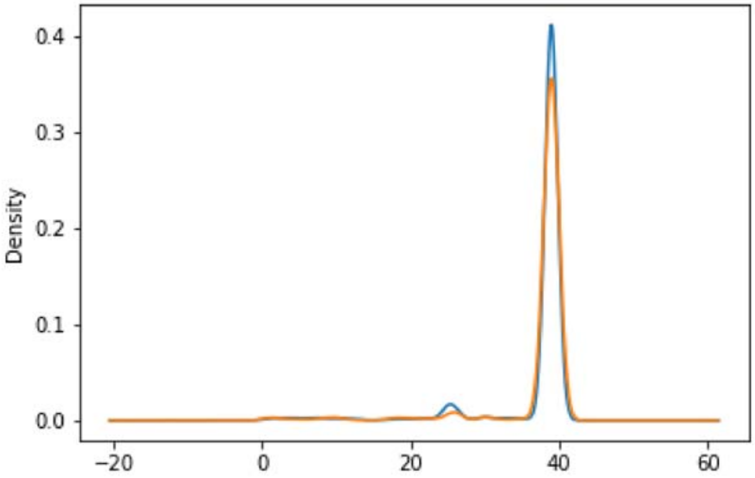
```
<matplotlib.figure.Figure at 0x7f6562486f28>
```

Native Country Histogram

In [34]:

```
income.groupby("target")["native-country"].plot(kind="kde")

plt.figure()
plt.show()
```



```
<matplotlib.figure.Figure at 0x7f6561c61cc0>
```

Conclusion

Drop column : fnlwgt, race, native-country karena tiap isi dari data tersebut tidak dapat menunjukkan data dominan pada range tertentu

misal : pada race kategori 4, data tidak dapat dibedakan antara kategori <=50k dan >50k