Analisis Dataset Adults (diambil dari UCI Machine Learning Repository)

Arief Akbar Hidayat Big Data B

Analisis dilakukan terhadap Adult data set yang diambil dari UCI Machine Learning Repository. Analisis ini dilakukan dengan algoritma knn untuk memprediksi apakah income seseorang melebihi 50K atau tidak per tahun berdasarkan dataset yang ada.

Eksplorasi data:

1. Menampilkan raw data terlebih dahulu. Dapat dilihat bahwa data terdiri atas 15 *rows* (baris) yang isi datanya terdapat variable kategorik dan variable numerik

```
print(df.head())
```

```
   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

       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

   capital-gain  capital-loss  hours-per-week native-country salary
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
```

2. Mengecek missing values data. Dapat dilihat bahwa tidak ada missing values yang terdapat dalam data ini.

```
df.isnull().sum()
```

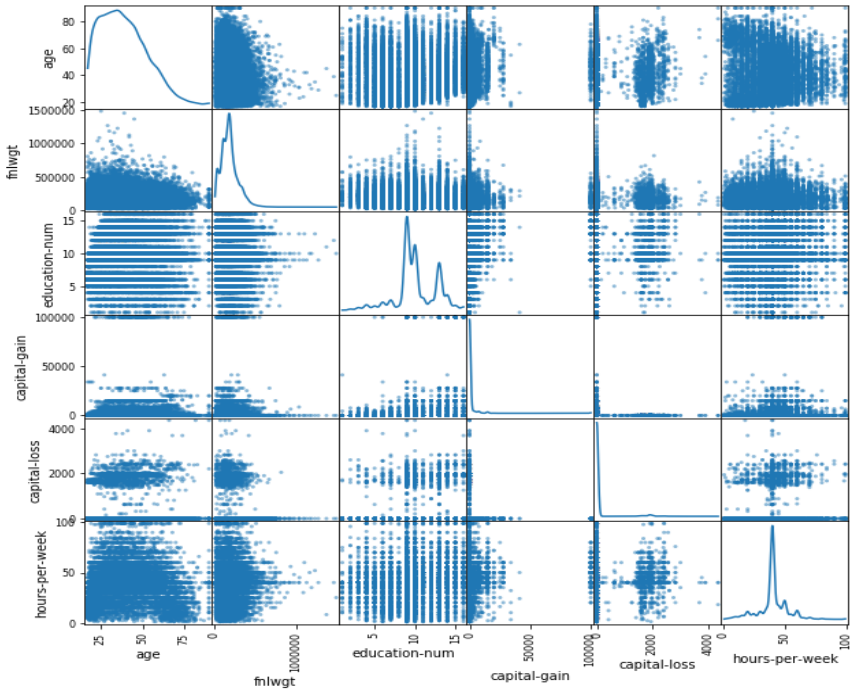
```
age               0
workclass         0
fnlwgt            0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
salary            0
dtype: int64
```

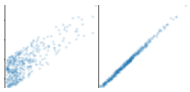
3. Analisis statistik untuk variable numerik yang terdapat dalam data.

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

Jumlah data keseluruhan adalah 32561 data. Rata-rata *age* dalam data tersebut berkisar di angka 38.58 tahun dan standar deviasi 13.64 yang memiliki arti bahwa data terdistribusi dan menyebar dengan cukup baik. Begitu pula halnya dengan kolom fnlwgt,education-num dan hours per-week yang memiliki penyebaran data dan distribusi yang cukup baik. Tetapi untuk capital-gain dan capital-loss terdapat kesenjangan yang tidak wajar antara rata-rata dengan standar deviasinya, lalu dapat dilihat juga ketiga kuartil pada dua kolom ini memiliki nilai nol sehingga menyebabkan penyebaran data dan distribusinya menjadi tidak normal.

4. Melakukan Bivariate Analysis
    a. Analisis variabel numerik dengan numerik
        1. membuat scatter plot untuk semua kolom yang memiliki variabel numerik



Dapat dilihat bahwa tidak ada titik-titk yang membentuk garis linear positif seperti ini 

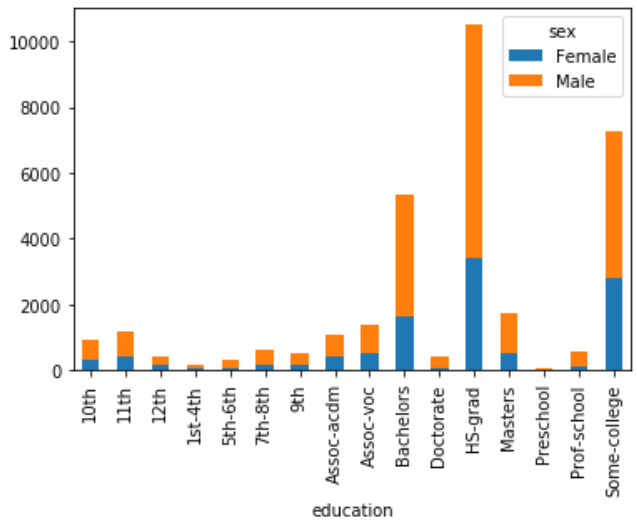
Maka dapat dipastikan bahwa tidak ada korelasi pada hubungan antar kolom yang memiliki variabel numerik.

Selanjutnya Mari kita analisis nilai koefisien r setiap kolom untuk melihat apakah ada korelasi yang kuat atau tidak.

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
age	1.000000	-0.076646	0.036527	0.077674	0.057775	0.068756
fnlwgt	-0.076646	1.000000	-0.043195	0.000432	-0.010252	-0.018768
education-num	0.036527	-0.043195	1.000000	0.122630	0.079923	0.148123
capital-gain	0.077674	0.000432	0.122630	1.000000	-0.031615	0.078409
capital-loss	0.057775	-0.010252	0.079923	-0.031615	1.000000	0.054256
hours-per-week	0.068756	-0.018768	0.148123	0.078409	0.054256	1.000000

Dapat kita lihat pada tabel korelasi, tidak ada nilai yang cukup besar (>0,6) pada hubungan setiap data numerik dengan data numerik yang lainya. Hal ini membuktikan bahwa data numerik dari dataset salary ini tidak saling berkolerasi, sehingga analisis selanjutnya ialah menggunakan bivariate analysis kategorik dan kategorik.

b. Analisis variabel kategorik dengan kategorik
    1) Analisis kolom kategorik sex dengan education



Two-way table antar variabel kolom sex dan education

sex	Female	Male	All
education			
10th	295	638	933
11th	432	743	1175
12th	144	289	433
1st-4th	46	122	168
5th-6th	84	249	333
7th-8th	160	486	646
9th	144	370	514
Assoc-acdm	421	646	1067
Assoc-voc	500	882	1382
Bachelors	1619	3736	5355
Doctorate	86	327	413
HS-grad	3390	7111	10501
Masters	536	1187	1723
Preschool	16	35	51
Prof-school	92	484	576
Some-college	2806	4485	7291
All	10771	21790	32561

Stack column chart dari two way variabel sex dan education

Kemudian lakukan perhitungan di python untuk mencari chi-square, p-value, critical-value dan degrees of freedom. Hasil yang didapat:

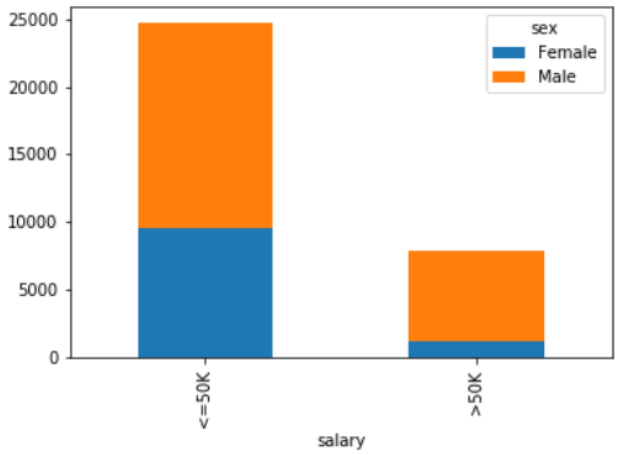
Chi-square=297.71500372503687,
P-value= 1.667778440920507e-54,
Critical-value= 3.841458820694124
Degrees of freedom = 15

Dapat dilihat bahwa nilai chi-square > critical value, yang memiliki arti kedua variabel antara kolom sex dengan kolom education saling berkorelasi (dependent).

2) Analisis kolom kategorik sex dengan salary

sex salary	Female	Male	All
<=50K	9592	15128	24720
>50K	1179	6662	7841
All	10771	21790	32561

Two-way table antar variabel kolom sex dan salary



Stack column chart dari two way variabel sex dan salary

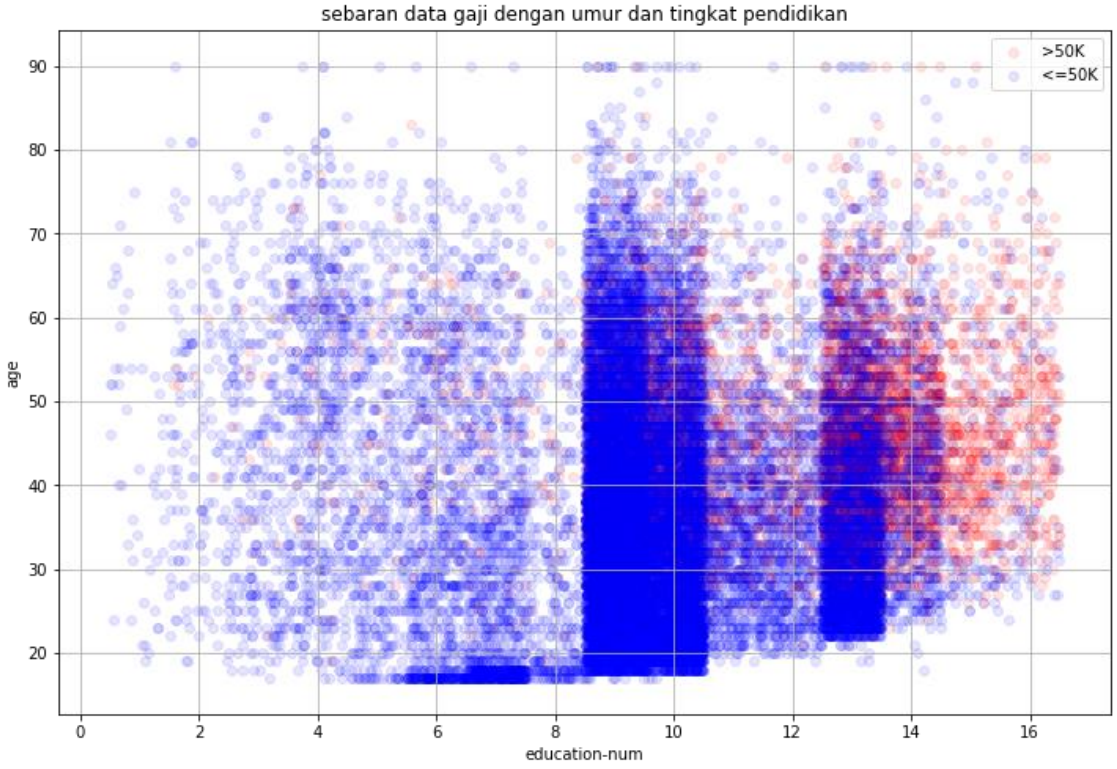
Kemudian lakukan perhitungan di python untuk mencari chi-square, p-value, critical-value dan degrees of freedom. Hasil yang didapat

Chi-square=1517.813409134445
P-value= 0.0
Critical-value= 3.841458820694124
Degrees of freedom = 1

Dapat dilihat bahwa nilai chi-square > critical value, yang memiliki arti kedua variabel antara kolom sex dengan kolom education saling berkorelasi (dependent).

Prediksi income seseorang melebihi 50K atau tidak per tahun berdasarkan dataset yang ada.

Sebelum melakukan prediksi, mari kita lihat sebaran data gaji,umur dan tingkat pendidikan seseorang

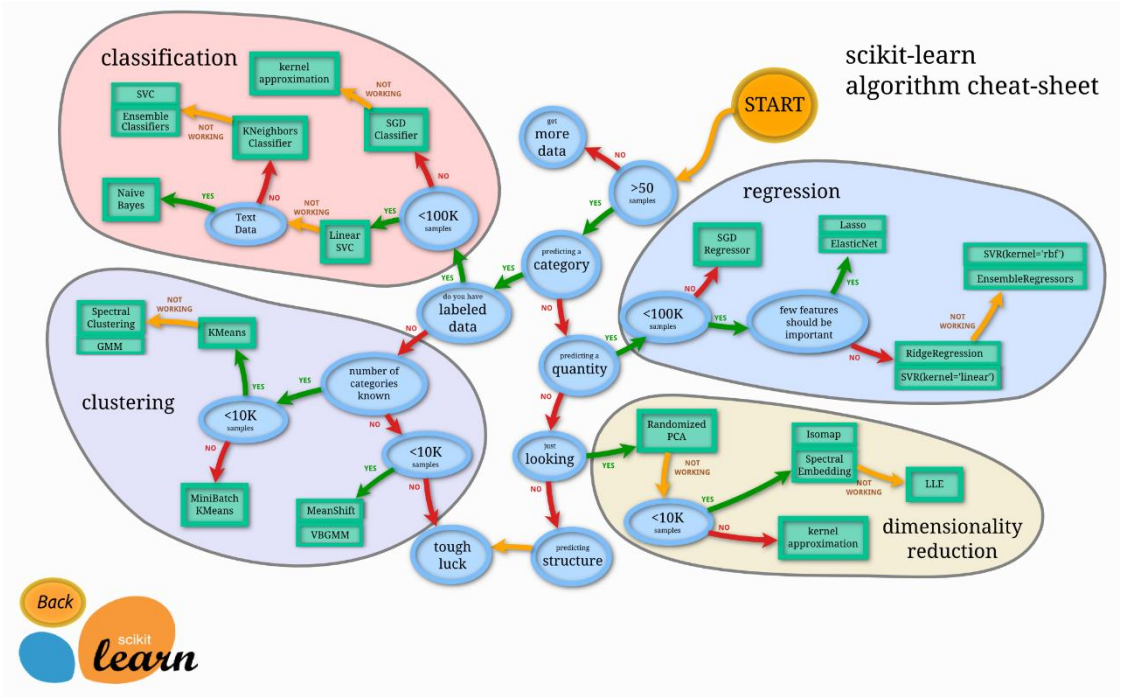


education-num	education
1	Preschool
2	1st-4th
3	5th-6th
4	7th-8th
5	9th
6	10th
7	11th
8	12th
9	HS-grad
10	Some-college
11	Assoc-voc
12	Assoc-acdm
13	Bachelors
14	Masters
15	Prof-school
16	Doctorate

Dari visualisasi tersebut dapat dilihat sebaran orang yang memiliki gaji <=50K berkumpul paling banyak pada education-num range 9-10. Sedangkan sebaran orang yang memiliki gaji >50K menyebar dan berkumpul di range 14-16. Penjelasan nomor education-num sudah tertera pada tabel.

Langkah-langkah metode prediksi income seseorang melebihi 50K atau tidak per tahun berdasarkan dataset yang ada dengan machine learning.

1. Pemilihan algoritma yang sesuai untuk prediksi data



Saya memilih path algoritma klasifikasi menggunakan kNeigbors classifier karena menurut saya itu yang paling sesuai dengan dataset yang ada.

2. Pemilihan data

Selanjutnya saya melakukan pemilihan data yang akan dipakai untuk keperluan machine learning. Dalam kasus ini saya membuang data capital-gain,capital-loss dikarenakan data tersebut tidak memiliki data yang cukup,karena sebagian besar data bernilai nol dan memang tidak diperlukan dalam pembuatan machine learning untuk memprediksi income seseorang. Lalu saya juga membuang salary, karena hasil prediksi yang akan dibuat adalah prediksi salary(income) itu sendiri. Kemudian saya juga membuang fnlwgt karena data tersebut memiliki range yang sangat besar dibanding dengan data lainnya sehingga dapat menyebabkan pembuatan machine learning yang tidak optimal.

3. Merubah data kategorik menjadi numerik, Saya menggunakan dictVectorizer untuk merubah data kategorik ke numerik. Setelah itu melakukan training data. Disini gaji >50K akan menjadi resultnya.

```
from sklearn.feature_extraction import DictVectorizer
vec= DictVectorizer()
features=vec.fit_transform(
                    df.dropna()\
                    .drop('capital-gain', axis=1)\
                    .drop('capital-loss', axis=1)\
                    .drop('salary', axis=1)\
                    .drop('fnlwgt', axis=1)\
                    .to_dict(orient='records')).toarray()

result=df['salary']=='>50K'
X_train=features[1000:]
Y_train=result[1000:]
X_test=features[:1000]
Y_test=result[:1000]
```

4. Mengetes hasil prediksinya

Yang perlu diperhatikan ada nilai recallnya. Karena semakin tinggi recall valuesnya maka akurasinya akan semakin tinggi juga. Dari sekian report yang didapat, saya memilih report yang memiliki recall dan akurasi tertinggi.

```
Accuracy: 83.0%, Neighbors: 13, weights: uniform
              precision    recall  f1-score   support

       <=50K       0.88      0.90      0.89       768
        >50K       0.65      0.61      0.63       232

   micro avg       0.83      0.83      0.83      1000
   macro avg       0.77      0.76      0.76      1000
weighted avg       0.83      0.83      0.83      1000
```

5. uji coba dengan input data

In [83]:
```
person={
    'age':40,
    'education':'Doctorate',
    'education-num':13,
    'hours-per-week':13,
    'marital-status':'Never-married',
    'native-country':'United-States',
    'occupation':'Prof-specialty',
    'race':'White',
    'relationship':'Husband',
    'sex':'Male',
    'workclass':'private'
}
person_features=vec.transform(person).toarray()
prediction=knn.predict(person_features)[0]
labels={True:'diatas 50 K', False:'dibawah 50 K'}
print(labels[prediction])
```

```
diatas 50 K
```