

LAPORAN TUGAS 2 *DATA MINING*



Disusun oleh:

Muhammad Naufal Syawali Akbar (1301164488)

Anasya Wulandari (1301174028)

Imam Nurul Ihsan (1301174688)

Program Studi Sarjana Informatika

Fakultas Informatika

Universitas Telkom

Bandung

2020

A. Latar Belakang Masalah

Data mining merupakan proses *iterative* untuk menemukan sesuatu dengan cara *automatic* atau manual. *Data mining* berfokus pada cara mengekstraksi *knowledge* atau pengetahuan yang menarik dalam bentuk data yang tersimpan di *database* yang berukuran besar. Selain itu, tujuan lainnya yaitu untuk mendapatkan *insight* dan sebagai cara dalam proses pengambilan keputusan. Data yang dihasilkan implisit, sebelumnya tidak diketahui, dan berpotensi berharga.

Task dalam *data mining* ada dua, yaitu prediksi dan deskripsi. Prediksi merupakan *task* dalam *data mining* yang menggunakan beberapa variabel untuk mengetahui atau memprediksi nilai yang belum diketahui. Dengan hal tersebut, fungsionalitas dari *data mining* sangatlah luas. Tidak terkecuali dengan klasifikasi sebuah data dan memprediksi keakuratannya. Salah satu contoh klasifikasi dalam *data mining* berada dalam dunia ekonomi dan kesehatan. Dalam dunia ekonomi *data mining* memprediksi dan klasifikasi pendapatan atau gaji, berbeda dengan dunia kesehatan *data mining* dapat memprediksi/mendiagnosa dan klasifikasi data penyakit seperti kanker, diabetes, dan yang lainnya.

Biasanya, untuk membangun sistem prediksi dan klasifikasi itu dibutuhkan waktu dan biaya yang tidak murah. Oleh karena itu, peran dalam *data mining* dirasa cocok untuk memangkas pengeluaran biaya dan lama pengerjaan dalam memprediksi dan klasifikasi sebuah data.

B. Tujuan

Tujuan dari penelitian ini diantaranya:

- Membangun sistem klasifikasi dan prediksi untuk *dataset* yang dipilih dengan metode *Naive Bayes* dan *Decision Tree*
- Mendapatkan akurasi dari prediksi dengan baik.
- Hasil dari penerapan *data mining* dalam sistem dapat digunakan dalam bidang yang bersangkutan sebagai informasi untuk digunakan di kemudian hari.
- Algoritma dengan tingkat akurasi yang baik dapat digunakan untuk penelitian klasifikasi selanjutnya.

C. Deskripsi Data

Dari 3 *dataset* yang disediakan, kelompok kami memilih 2 *dataset*. Diantaranya:

- <https://archive.ics.uci.edu/ml/datasets/Census+Income>
- <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer>

Dataset tersebut bersumber dari UCI *machine learning repository*. Untuk *dataset Census Income* sendiri karakteristik *dataset* nya yaitu *multivariate*, karakteristik atributnya *categorical* dan integer, dengan jumlah data sebanyak 48842 dengan atribut sebanyak 14. *Dataset* ini juga memiliki *missing value*. Sedangkan, Untuk *dataset Breast Cancer* sendiri karakteristik *dataset* nya yaitu *multivariate*, karakteristik atributnya *categorical*, dengan jumlah data sebanyak 286 dengan atribut sebanyak 9: class, age, menopause, tumor-size, inv-nodes, node-caps, deg-malig, breast, breast-quad, irradiate. *Dataset* ini juga memiliki *missing value* dan *wrong data*. Dengan adanya *missing value* di dalam kedua *dataset*, bisa dikatakan *dataset* tersebut belum berkualitas. Sehingga, diperlukannya praproses data untuk menghasilkan data yang berkualitas untuk diolah.

D. Preprocessing

1. Preprocessing Data Census

Preprocessing data adalah strategi dan teknik yang saling berkaitan untuk membuat data lebih mudah atau cocok untuk digunakan. *Preprocessing* bertujuan untuk menghasilkan meningkatkan kualitas data, juga hasil analisis data. Pada percobaan penelitian ini kelompok kami mencoba melakukan *preprocessing* seperti menghitung *missing values*, dan mengganti nilai *missing values* agar data menjadi berkualitas dan dapat diolah. Kondisi data setelah melakukan *preprocessing* dapat dibuktikan dalam gambar dibawah ini.

In [16]:

df.describe(include=["O"])

Out[16]:

	workclass	education	marital.status	occupation	relationship	race	sex	native.country	Income
count	32561	32561	32561	32561	32561	32561	32561	32561	32561
unique	9	16	7	15	6	5	2	42	2
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United States	<=50K
freq	22696	10501	14976	4140	13193	27816	21790	29170	24720

In [17]:

df.isnull().sum()

Out[17]:

age0
workclass0
fnlwgt0
education0
education.num0
marital.status0
occupation0
relationship0
race0
sex0
capital.gain0
capital.loss0
hours.per.week0
native.country0
income0
dtype: int64

In [18]:

df

Out[18]:

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.loss	hours.per.week	n
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	
4	28	Private	336409	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	
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	

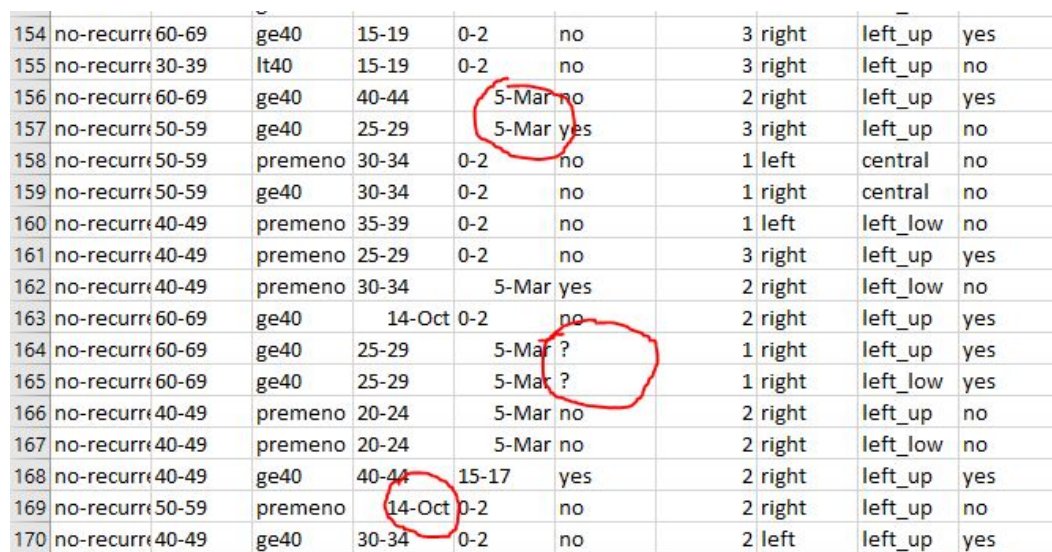
32561 rows x 15 columns

Gambar 1. Hasil preprocessing

Terlihat bahwa *dataset* tidak lagi terdapat *missing value*, Dengan begitu *dataset* bisa disebutkan sebagai data yang berkualitas yang nantinya akan digunakan untuk dilakukan klasifikasi dan prediksi untuk metode algoritma yang dipilih.

2. *Preprocessing Breast Cancer*

Karena dataset breast cancer merupakan dataset yang kualitasnya kurang dan untuk dapat diolah dataset harus dalam kondisi yang berkualitas, oleh karena itu kami melakukan beberapa metode seperti menghapus missing values, dan mengganti wrong data yang terdapat pada dataset breast cancer. Pada Gambar 2. lingkaran merah merupakan salah satu dari missing values, dan juga wrong data pada kolom “tumor-size” yang seharusnya merupakan kisaran dari ukuran tumor tetapi pada dataset tersebut adalah tanggal, ini merupakan contoh dari sekian banyak wrong data yang terdapat pada dataset. Untuk penyelesaian masalah missing values, kami memutuskan untuk menggunakan *Wise Deletion*. *Wise deletion* merupakan salah satu metode untuk menghapus missing values pada dataset, wise deletion bekerja dengan menghapus satu baris tabel yang terdapat missing values, cukup sederhana untuk digunakan namun metode wise deletion dapat mengurangi jumlah dataset, karena dalam dataset Breast Cancer ini tidak terlalu banyak missing values, maka wise deletion dapat dibilang efektif.



154	no-recurrence-events	60-69	ge40	15-19	0-2	no	3	right	left_up	yes
155	no-recurrence-events	30-39	lt40	15-19	0-2	no	3	right	left_up	no
156	no-recurrence-events	60-69	ge40	40-44	5-Mar	no	2	right	left_up	yes
157	no-recurrence-events	50-59	ge40	25-29	5-Mar	yes	3	right	left_up	no
158	no-recurrence-events	50-59	premeno	30-34	0-2	no	1	left	central	no
159	no-recurrence-events	50-59	ge40	30-34	0-2	no	1	right	central	no
160	no-recurrence-events	40-49	premeno	35-39	0-2	no	1	left	left_low	no
161	no-recurrence-events	40-49	premeno	25-29	0-2	no	3	right	left_up	yes
162	no-recurrence-events	40-49	premeno	30-34	5-Mar	yes	2	right	left_low	no
163	no-recurrence-events	60-69	ge40	14-Oct	0-2	no	2	right	left_up	yes
164	no-recurrence-events	60-69	ge40	25-29	5-Mar	?	1	right	left_up	yes
165	no-recurrence-events	60-69	ge40	25-29	5-Mar	?	1	right	left_low	yes
166	no-recurrence-events	40-49	premeno	20-24	5-Mar	no	2	right	left_up	no
167	no-recurrence-events	40-49	premeno	20-24	5-Mar	no	2	right	left_low	no
168	no-recurrence-events	40-49	ge40	40-44	15-17	yes	2	right	left_up	yes
169	no-recurrence-events	50-59	premeno	14-Oct	0-2	no	2	right	left_up	no
170	no-recurrence-events	40-49	ge40	30-34	0-2	no	2	left	left_up	yes

Gambar 2. Missing values dan Wrong data pada dataset breast cancer

Pada dataset breast cancer terdapat banyak wrong data, seperti data tanggal yang terdapat pada kolom tumor size, karena data yang terdapat pada kolom tumor size adalah data dengan tipe String yang berisi perkiraan ukuran tumor, maka data dengan tipe Date dianggap adalah wrong data (salah data). Untuk kasus ini kami menyimpulkan untuk menggunakan data mayoritas yang terdapat pada ukuran tumor atau tumor size, maka dari itu data yang bertipe Date tersebut akan diganti dengan ukuran tumor “30-34” karena ukuran tersebut adalah ukuran yang paling sering muncul pada dataset *Breast Cancer*.

Setelah kami menghilangkan missing values dan mengganti wrong data pada dataset *Breast Cancer*. Karena pada tahap klasifikasi tidak dapat menggunakan data string maka langkah selanjutnya adalah mentransformasi atau encode terhadap label klasifikasi pada kolom “class” seperti: “no-recurrence-events” menjadi variabel

integer bernilai “0” dan “recurrence-events” menjadi variabel integer yang bernilai “1”, untuk ilustrasi dapat dilihat pada gambar 3(a) dan 3(b).

	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	no-recurrence-events	30-39	premeno	30-34	0-2	no	3	left	left_low	no
2	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	right	right_up	no
3	no-recurrence-events	40-49	premeno	20-24	0-2	no	2	left	left_low	no
4	no-recurrence-events	60-69	ge40	15-19	0-2	no	2	right	left_up	no
...
273	recurrence-events	30-39	premeno	30-34	0-2	no	2	left	left_up	no
274	recurrence-events	30-39	premeno	20-24	0-2	no	3	left	left_up	yes
275	recurrence-events	60-69	ge40	20-24	0-2	no	1	right	left_up	no
276	recurrence-events	40-49	ge40	30-34	0-2	no	3	left	left_low	no
277	recurrence-events	50-59	ge40	30-34	0-2	no	3	left	left_low	no

278 rows x 10 columns

Gambar 3(a). Dataset sebelum di transformasi atau encode

```
df['class'] = df['class'].map({'no-recurrence-events':0, 'recurrence-events':1})
df.head()
```

	class	age	menopause	tumor-size	inv-nodes	node-caps	deg-malig	breast	breast-quad	irradiat
0	0	30-39	premeno	30-34	0-2	no	3	left	left_low	no
1	0	30-39	premeno	30-34	0-2	no	3	left	left_low	no
2	0	40-49	premeno	20-24	0-2	no	2	right	right_up	no
3	0	40-49	premeno	20-24	0-2	no	2	left	left_low	no
4	0	60-69	ge40	15-19	0-2	no	2	right	left_up	no

Gambar 3(b). Dataset setelah di transformasi atau encode

E. Analisis Pemilihan Algoritma

1. Dataset Census Income

Pada *dataset* ini penulis memilih 2 algoritma untuk melakukan klasifikasi. Algoritma tersebut adalah *Naive Bayes* dan *Decision Tree*. Kedua algoritma tersebut digunakan untuk membandingkan akurasi terbaik. Salah satu alasan menggunakan algoritma *Naive Bayes* adalah karena bisa dipakai untuk data kuantitatif maupun kualitatif. Sedangkan algoritma *Decision Tree* dipilih karena seringkali memiliki nilai akurasi yang lebih baik.

2. Dataset Breast Cancer

Pada *dataset Breast Cancer*, penulis memilih *Gaussian Naive Bayes* dan *Decision Tree*. Seperti pada dataset *Census Income* alasan menggunakan kedua metode classifier adalah untuk membandingkan model mana yang terbaik dalam mengklasifikasi pada kasus dataset *Breast Cancer*. Naive Bayes merupakan pengklasifikasi berdasarkan asumsi keberadaan fitur tertentu dalam suatu kelas tidak akan terkait dengan keberadaan fitur lainnya. sedangkan *Decision Tree* memiliki kinerja yang mirip dengan Random Forest yaitu membuat keputusan akhir berdasarkan keputusan - keputusan yang berada pada semua pohon.

F. Analisis Penentuan Parameter

1. Dataset Census Income

Parameter yang digunakan penulis dalam *dataset* ini diantaranya pengisian *missing value* dengan nilai yang paling sering muncul pada setiap kolomnya, *drop column* dilakukan ketika kolom mendapatkan korelasi 0, memisahkan kumpulan data menjadi fitur dan hasil, dan terakhir membagi data menjadi data uji dan data *test* dengan ukuran *test* 0.3.

2. Dataset Breast Cancer

Parameter yang digunakan penulis dalam *dataset* ini diantaranya penghapusan *missing values* dengan metode *wise deletion*, pengisian *wrong data* dengan nilai yang paling sering muncul pada kolom “tumor-size”, dan encode label pada kolom “class” dari string menjadi nilai integer seperti: “no-recurrence-events” menjadi variabel integer bernilai “0” dan “recurrence-events” menjadi variabel integer yang bernilai “1”. Sebelum menggunakan classifier, perlu dilakukan pembagian dataset terlebih dahulu, untuk tahap training dan tahap testing. Pembagian untuk tahap training dan testing secara berurutan sebesar 3:1 dengan ukuran “test-size” sebesar 0.25. Tujuannya adalah menambahkan data pada tahap training agar model memiliki kualitas dan banyak kesamaan fitur sehingga akan meningkatkan nilai akurasi pada tahap testing.

G. Hasil Percobaan

1. Dataset Census Income

Hasil yang diperoleh dalam penelitian *dataset* ini adalah sebagai berikut:

```
In [32]: # Akurasi

results = pd.DataFrame({
    'Model': ['Decision Tree', 'Naive Bayes'],
    'Score': [acc_decision_tree, acc_gaussian]})
result_df = results.sort_values(by='Score', ascending=False)
result_df = result_df.set_index('Score')
result_df
```

Out[32]:

	Model
Score	
98.01	Decision Tree
80.29	Naive Bayes

Gambar 4. Hasil Akurasi Dataset 1

Dapat dilihat bahwa algoritma *decision tree* memperoleh hasil akurasi yang lebih baik dari algoritma *naive bayes* yaitu sebesar 98,01%. Hal tersebut sesuai dengan penjelasan pada bagian E bahwa, algoritma *decision tree* memang seringkali mendapatkan hasil nilai akurasi yang baik. Sesuai dengan tujuan pada poin B, maka algoritma *decision tree* dapat digunakan untuk penelitian klasifikasi selanjutnya.

2. Dataset Breast Cancer

Hasil yang diperoleh dalam penelitian dataset ini adalah sebagai berikut:

```
#Using DecisionTreeClassifier of tree class to use Decision Tree Algorithm

from sklearn.tree import DecisionTreeClassifier
classifier = DecisionTreeClassifier(criterion = 'entropy')
classifier.fit(x_train, y_train)

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                        max_depth=None, max_features=None, max_leaf_nodes=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, presort='deprecated',
                        random_state=20, splitter='best')

[40] from sklearn.metrics import accuracy_score
y_pred = classifier.predict(x_test)
print("Akurasi", accuracy_score(y_test, y_pred))

Akurasi 0.6857142857142857
```

Gambar 5 (a). Hasil Akurasi dari Decision Tree

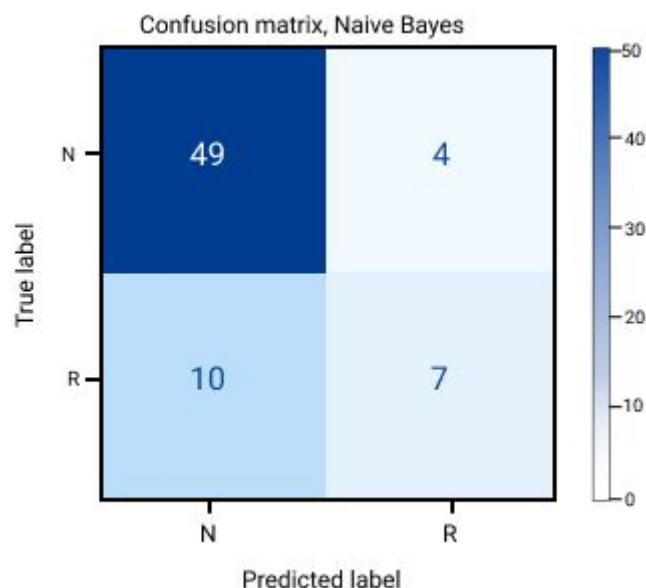
```
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()
classifier.fit(x_train, y_train)

GaussianNB(priors=None, var_smoothing=1e-09)

from sklearn.metrics import accuracy_score
y_pred = classifier.predict(x_test)
print("Akurasi", accuracy_score(y_test, y_pred))

Akurasi 0.8
```

Gambar 5 (b). Hasil Akurasi dari Naive Bayes



Gambar 5 (c). Naive Bayes Confusion Matrix

Dari gambar diatas dapat dilihat bahwa *Naive Bayes* memiliki performansi yang lebih baik daripada *Decision Tree* dengan akurasi sebesar 80% sedangkan akurasi *Decision Tree* sebesar 68.5%, kemungkinan jumlah dataset yang sedikit membuat akurasi *Decision Tree* berkurang, karena jumlah random state pohon yang sedikit. Dari gambar 5(c), terdapat 49 data testing no-recurrence positif. dan 4 no-recurrence terprediksi sebagai recurrence. pada baris kedua terdapat 10 recurrence yang terprediksi no-recurrence dan terakhir ada 7 data recurrence terprediksi positif recurrence.

H. Ringkasan Model Yang Diperoleh

Buat ringkasan model yang anda pilih (jika memang tidak semua model dari hasil tahap G digunakan).

1. Dataset Census Income

Dalam menyelesaikan permasalahan yang ada digunakan metode klasifikasi yaitu menggunakan Decision Tree. Pemilihan data dilakukan dengan menggunakan library, dan atribut yang berbentuk categorical diubah menggunakan labelencoder, dan normalisasi data.

2. Dataset Breast Cancer

dalam menyelesaikan klasifikasi akhir, metode yang mendapat hasil terbaik adalah Naive Bayes, pemilihan data dalam tahap pemisahan data training dan testing menggunakan library dari sklearn begitupun dalam tahap klasifikasi menggunakan library dari sklearn untuk *Naive Bayes*, Decision Tree dan *Confusion Matrix*.

I. Interpretasi Model

Jelaskan makna dari model yang anda peroleh, tunjukkan apakah tujuan yang telah disebutkan pada bagian B tercapai atau tidak. Jika tercapai/tidak jelaskan alasannya.

Jelaskan pula apakah dengan model yang diperoleh masalah yang telah disebutkan di bagian A benar-benar telah didapat solusinya atau tidak.

1. Dataset Census Income

Model yang diperoleh telah memenuhi tujuan yang disebutkan pada bagian B. Hal ini dikarenakan terdapat hasil klasifikasi dengan akurasi yang tinggi dengan menggunakan metode *Decision Tree*. Pada proses klasifikasi terdapat data yang tidak digunakan, karena korelasi yang bernilai 0 dan tidak adanya keterkaitan data tersebut dengan parameter yang telah dipilih.

2. Dataset Breast Cancer

Model yang diperoleh telah memenuhi tujuan yang disebutkan pada bagian B. hal ini dikarenakan terdapat hasil performansi klasifikasi dengan menggunakan metode Naive Bayes.