Klasifikasi dengan menggunakan Algoritma Naïve Bayes

Disusun oleh :

Mila Putri Kartika Dewi (1301174218)

Dzaka Triadi Mahdiyah (1301152728)

Wilrades Christofel (1301170753)

PROGRAM STUDI S1 INFORMATIKA

FAKULTAS INFORMATIKA

TELKOM UNIVERSITY

BANDUNG

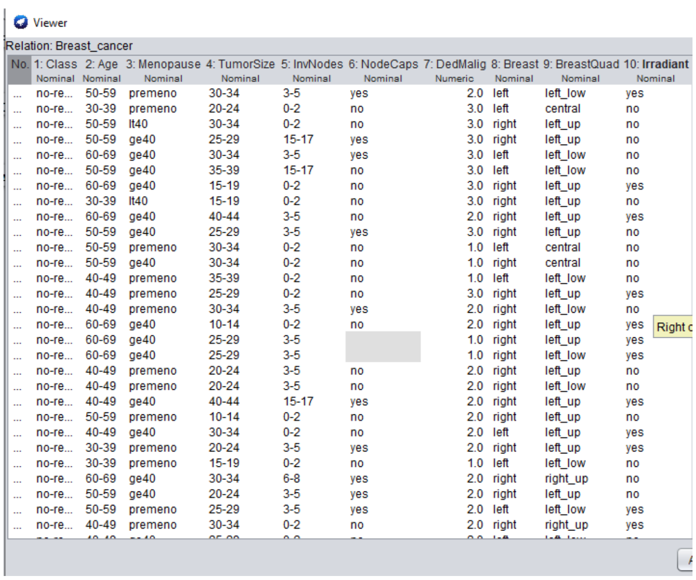
2020

1. Dataset
   Dalam tugas analisi ini, kelompok kami menggunakan dataset dari UCI Machine Learning Repository, dataset yang kami gunakan adalah :
   - Breast Cancer
   - Cencus Income

2. Analisis Data (Pre-Processing)

   2.1 Breast Cancer
   Dataset Breast Cancer memiliki record sebanyak 286, namun data ini belum terdapat atribut, sehingga harus ditambahkan secara manual. Data Breast Cancer dikatakan tidak layak karena masih terdapat missing value, sehingga perlu dilakukan pre-processing dari data tersebut. Berikut langkah-langkah untuk menangani masalah missing value :
   - Melakukan filter data dengan memilih **Filters** > **Unsupervised** > **Attribute** > **ReplaceMissingValues.**

   - Data akan otomatis terisikan denga nilai rata-rata dari seriap atribut tersebut, berikut perbandingan nilai sebelum dan sesudah dilakukannya replace missing value :


Viewer

Relation: Breast_cancer

No.	1: Class Nominal	2: Age Nominal	3: Menopause Nominal	4: TumorSize Nominal	5: InvNodes Nominal	6: NodeCaps Nominal	7: DedMalig Numeric	8: Breast Nominal	9: BreastQuad Nominal	10: Irradiant Nominal
...	no-re...	50-59	premeno	30-34	3-5	yes	2.0	left	left_low	yes
...	no-re...	30-39	premeno	20-24	0-2	no	3.0	left	central	no
...	no-re...	50-59	lt40	30-34	0-2	no	3.0	right	left_up	no
...	no-re...	50-59	ge40	25-29	15-17	yes	3.0	right	left_up	no
...	no-re...	60-69	ge40	30-34	3-5	yes	3.0	left	left_low	no
...	no-re...	50-59	ge40	35-39	15-17	no	3.0	left	left_low	no
...	no-re...	60-69	ge40	15-19	0-2	no	3.0	right	left_up	yes
...	no-re...	30-39	lt40	15-19	0-2	no	3.0	right	left_up	no
...	no-re...	60-69	ge40	40-44	3-5	no	2.0	right	left_up	yes
...	no-re...	50-59	ge40	25-29	3-5	yes	3.0	right	left_up	no
...	no-re...	50-59	premeno	30-34	0-2	no	1.0	left	central	no
...	no-re...	50-59	ge40	30-34	0-2	no	1.0	right	central	no
...	no-re...	40-49	premeno	35-39	0-2	no	1.0	left	left_low	no
...	no-re...	40-49	premeno	25-29	0-2	no	3.0	right	left_up	yes
...	no-re...	40-49	premeno	30-34	3-5	yes	2.0	right	left_low	no
...	no-re...	60-69	ge40	10-14	0-2	no	2.0	right	left_up	yes
...	no-re...	60-69	ge40	25-29	3-5		1.0	right	left_up	yes
...	no-re...	60-69	ge40	25-29	3-5		1.0	right	left_low	yes
...	no-re...	40-49	premeno	20-24	3-5	no	2.0	right	left_up	no
...	no-re...	40-49	premeno	20-24	3-5	no	2.0	right	left_low	no
...	no-re...	40-49	ge40	40-44	15-17	yes	2.0	right	left_up	yes
...	no-re...	50-59	premeno	10-14	0-2	no	2.0	right	left_up	no
...	no-re...	40-49	ge40	30-34	0-2	no	2.0	left	left_up	yes
...	no-re...	30-39	premeno	20-24	3-5	yes	2.0	right	left_up	yes
...	no-re...	30-39	premeno	15-19	0-2	no	1.0	left	left_low	no
...	no-re...	60-69	ge40	30-34	6-8	yes	2.0	right	right_up	no
...	no-re...	50-59	ge40	20-24	3-5	yes	2.0	right	left_up	no
...	no-re...	50-59	premeno	25-29	3-5	yes	2.0	left	left_low	yes
...	no-re...	40-49	premeno	30-34	0-2	no	2.0	right	right_up	yes

Gambar 1. sebelum pre-processing

Viewer

Relation: Breast_cancer-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: Class Nominal	2: Age Nominal	3: Menopause Nominal	4: TumorSize Nominal	5: InvNodes Nominal	6: NodeCaps Nominal	7: DedMalig Numeric	8: Breast Nominal	9: BreastQuad Nominal	10: Irradiant Nominal
...	no-re...	30-39	premeno	15-19	0-2	no	1.0	left	left_low	no
...	no-re...	60-69	ge40	30-34	6-8	yes	2.0	right	right_up	no
...	no-re...	50-59	ge40	20-24	3-5	yes	2.0	right	left_up	no
...	no-re...	50-59	premeno	25-29	3-5	yes	2.0	left	left_low	yes
...	no-re...	40-49	premeno	30-34	0-2	no	2.0	right	right_up	yes
...	no-re...	40-49	ge40	25-29	0-2	no	2.0	left	left_low	no
...	no-re...	60-69	ge40	10-14	0-2	no	2.0	left	left_low	no
...	no-re...	50-59	premeno	25-29	3-5	no	2.0	right	left_up	yes
...	no-re...	40-49	premeno	20-24	0-2	no	3.0	right	left_low	yes
...	no-re...	40-49	premeno	35-39	0-2	yes	3.0	right	left_up	yes
...	no-re...	40-49	premeno	35-39	0-2	yes	3.0	right	left_low	yes
...	no-re...	40-49	premeno	25-29	0-2	no	1.0	right	left_low	yes
...	no-re...	50-59	ge40	30-34	9-11	no	3.0	left	left_up	yes
...	no-re...	50-59	ge40	30-34	9-11	no	3.0	left	left_low	yes
...	no-re...	40-49	premeno	20-24	6-8	no	2.0	right	left_low	yes
...	no-re...	50-59	ge40	25-29	0-2	no	1.0	left	right_low	no
...	no-re...	60-69	ge40	15-19	0-2	no	2.0	left	left_up	yes
...	no-re...	40-49	premeno	10-14	0-2	no	2.0	right	left_up	no
...	no-re...	50-59	ge40	20-24	0-2	yes	2.0	right	left_up	no
...	no-re...	40-49	premeno	15-19	12-14	no	3.0	right	right_low	yes
...	no-re...	40-49	premeno	25-29	0-2	no	2.0	left	left_up	yes
...	no-re...	50-59	ge40	30-34	6-8	yes	2.0	left	left_low	no
...	no-re...	30-39	premeno	10-14	0-2	no	2.0	left	right_low	no
...	no-re...	50-59	premeno	50-54	0-2	yes	2.0	right	left_up	yes
...	no-re...	50-59	ge40	35-39	0-2	no	2.0	left	left_up	no
...	no-re...	50-59	premeno	10-14	3-5	no	1.0	right	left_up	no
...	no-re...	40-49	premeno	10-14	0-2	no	2.0	left	left_low	yes
...	no-re...	50-59	ge40	15-19	0-2	yes	2.0	left	central	yes
...	no-re...	50-59	premeno	25-29	0-2	no	1.0	left	left_low	no

Gambar 2. Sesudah pre-precessing

- Label Encoding

  Data yang akan diproses harus dirubah kedalam bentuk numeric. Seperti nilai atribut “Menapause” yang bertype string kemudian dilakukan proses encoding dengan nilai dari “premeno” = 2, “g40” = 0 dan “it40” = 1. Dalam pre processing tahap ini, dilakukan encoding label dengan menggunakan algoritma python dan *LabelEncoder*. Berikut penjelasan algoritma dan tampilan atau hasil output dari encoding pada atribut “Menapause”:

```
labelencoder = LabelEncoder()
data['Class_trns'] = labelencoder.fit_transform(data['Class'])
data['Age_trans'] = labelencoder.fit_transform(data['Age'])
data['Menopause_trans'] = labelencoder.fit_transform(data['Menopause'])
data['TumorSize_trans'] = labelencoder.fit_transform(data['TumorSize'])
data['InvNodes_trans'] = labelencoder.fit_transform(data['InvNodes'])
data['NodeCaps_trans'] = labelencoder.fit_transform(data['NodeCaps'])
data['Breast_trans'] = labelencoder.fit_transform(data['Breast'])
data['BreastQuad_trans'] = labelencoder.fit_transform(data['BreastQuad'])
data['Irradiant_trans'] = labelencoder.fit_transform(data['Irradiant'])
```

Gambar 3. label encoding pada dataset breast cancer

Age_trans	TumorSize_trans	InvNodes_trans	NodeCaps_trans	Class_trns	Menopause_trans	Breast_trans	BreastQuad_trans
1	5	0	1	0	2	0	2
2	3	0	1	0	2	1	5
2	3	0	1	0	2	0	2
4	2	0	1	0	0	1	3
2	0	0	1	0	2	1	4

Gambar 4. hasil dari label encoding

2.2 Census Income

Dataset adult memiliki record sebanyak 32.561, data ini belum memiliki atribut sehingga perlu ditambahkan atribut baru untuk melakukan pre-processing. Nama atribut yang digunakan sesuai dengan sumber-sumber yang ada sebelumnya serta nama yang ditentukan juga sesuai dengan isi dari data tersebut, dan pada preprocessing dataset ini menggunakan perangkat lunak weka. Hasil analisis yang kelompok kami lakukan, dataset tersebut tidak berkualitas karena masih terdapat beberapa isi dari data yang kosong atau missing value, sehingga dataset tersebut hanya menampilakn nilai “Tanda Tanya” saja sehingga perlu dilakukannya pengisian pada data tersebut. Permasalahan yang terdapat pada data ini selain missing value juga terdapat data outlier, data ini cukup mencolok ketika atribut “Capital Gain” dan “Capital Loss” memiliki nilai yang sangat berbeda jauh, dengan isi 0.0 sebanyak 31.042 dan sisanya bernilai ratusan hingga ribuah, sehingga bisa dikatakan hampir keseluruhan data bernilai 0.0

Dengan permasalahan yang ada, maka perlu dilakukannya pre-processing pada dataset adult, menurut kelompok kami proses yang harus dilakukan seperti pengisian data yang kosong, dan penghapusan nilai yang outlier. Berikut penjelasan dari tahapan pre-processing data :

- Missing value

Dataset tersebut masih terdapat missing value, sehingga yang perlu dilakukan adalah mengisi nilai dari data tersebut dengan nilai baru. Cara kerja dari system ini yaitu mengisi nilai yang kosonh atau nilai yang hilang tersebut dengan nilai rata-rata dari nilai setiap atributnya, berikut langkah-langkah melakukan replace missing value :

1. Melakukan filter data dengan memilih **Filters** > **Unsupervised** > **Attribute** > **ReplaceMissingValues.**
2. Data akan otomatis terisikan denga nilai rata-rata dari seriap atribut tersebut, berikut perbandingan nilai sebelum dan sesudah dilakukannya replace missing value :

No.	1: Age	2: Workclass	3: Fnlwgt	4: Education	5: Education-Num	6: Matrial Status	7: Occupation	8: Relationship
	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal	Nominal
1	39	State-gov	77516	Bachelors	13	Never-married	Adm-cleric...	Not-in-family
2	50	Self-emp-...	83311	Bachelors	13	Married-civ-s...	Exec-man...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-...	Not-in-family
4	53	Private	234721	11th	7	Married-civ-s...	Handlers-...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-s...	Prof-speci...	Wife
6	37	Private	284582	Masters	14	Married-civ-s...	Exec-man...	Wife
7	49	Private	160187	9th	5	Married-spo...	Other-serv...	Not-in-family
8	52	Self-emp-...	209642	HS-grad	9	Married-civ-s...	Exec-man...	Husband
9	31	Private	45781	Masters	14	Never-married	Prof-speci...	Not-in-family
10	42	Private	159449	Bachelors	13	Married-civ-s...	Exec-man...	Husband
11	37	Private	280464	Some-co...	10	Married-civ-s...	Exec-man...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-s...	Prof-speci...	Husband
13	23	Private	122272	Bachelors	13	Never-married	Adm-cleric...	Own-child
14	32	Private	205019	Assoc-ac...	12	Never-married	Sales	Not-in-family
15	40	Private	121772	Assoc-voc	11	Married-civ-s...	Craft-repair	Husband
16	34	Private	245487	7th-8th	4	Married-civ-s...	Transport-...	Husband
17	25	Self-emp-...	176756	HS-grad	9	Never-married	Farming-fi...	Own-child
18	32	Private	186824	HS-grad	9	Never-married	Machine-o...	Unmarried
19	38	Private	28887	11th	7	Married-civ-s...	Sales	Husband
20	43	Self-emp-...	292175	Masters	14	Divorced	Exec-man...	Unmarried
21	40	Private	193524	Doctorate	16	Married-civ-s...	Prof-speci...	Husband
22	54	Private	302146	HS-grad	9	Separated	Other-serv...	Unmarried
23	35	Federal-g...	76845	9th	5	Married-civ-s...	Farming-fi...	Husband
24	43	Private	117037	11th	7	Married-civ-s...	Transport-...	Husband
25	59	Private	109015	HS-grad	9	Divorced	Tech-sup...	Unmarried
26	56	Local-gov	216851	Bachelors	13	Married-civ-s...	Tech-sup...	Husband
27	19	Private	168294	HS-grad	9	Never-married	Craft-repair	Own-child
28	54	?	180211	Some-co...	10	Married-civ-s...	?	Husband
29	39	Private	367260	HS-grad	9	Divorced	Exec-man...	Not-in-family
30	49	Private	193366	HS-grad	9	Married-civ-s...	Craft-repair	Husband

Gambar 5. sebelum pre-processing

No.	1: Age	2: Workclass	3: Fnlwgt	4: Education	5: Education-Num	6: Matrial Status	7: Occupation	8: Relationship
	Numeric	Nominal	Numeric	Nominal	Numeric	Nominal	Nominal	Nominal
1	39.0	State-gov	77516...	Bachelors	13.0	Never-married	Adm-cleric...	Not-in-family
2	50.0	Self-emp-...	83311...	Bachelors	13.0	Married-civ-s...	Exec-man...	Husband
3	38.0	Private	21564...	HS-grad	9.0	Divorced	Handlers-...	Not-in-family
4	53.0	Private	23472...	11th	7.0	Married-civ-s...	Handlers-...	Husband
5	28.0	Private	33840...	Bachelors	13.0	Married-civ-s...	Prof-speci...	Wife
6	37.0	Private	28458...	Masters	14.0	Married-civ-s...	Exec-man...	Wife
7	49.0	Private	16018...	9th	5.0	Married-spo...	Other-serv...	Not-in-family
8	52.0	Self-emp-...	20964...	HS-grad	9.0	Married-civ-s...	Exec-man...	Husband
9	31.0	Private	45781...	Masters	14.0	Never-married	Prof-speci...	Not-in-family
10	42.0	Private	15944...	Bachelors	13.0	Married-civ-s...	Exec-man...	Husband
11	37.0	Private	28046...	Some-co...	10.0	Married-civ-s...	Exec-man...	Husband
12	30.0	State-gov	14129...	Bachelors	13.0	Married-civ-s...	Prof-speci...	Husband
13	23.0	Private	12227...	Bachelors	13.0	Never-married	Adm-cleric...	Own-child
14	32.0	Private	20501...	Assoc-ac...	12.0	Never-married	Sales	Not-in-family
15	40.0	Private	12177...	Assoc-voc	11.0	Married-civ-s...	Craft-repair	Husband
16	34.0	Private	24548...	7th-8th	4.0	Married-civ-s...	Transport-...	Husband
17	25.0	Self-emp-...	17675...	HS-grad	9.0	Never-married	Farming-fi...	Own-child
18	32.0	Private	18682...	HS-grad	9.0	Never-married	Machine-o...	Unmarried
19	38.0	Private	28887...	11th	7.0	Married-civ-s...	Sales	Husband
20	43.0	Self-emp-...	29217...	Masters	14.0	Divorced	Exec-man...	Unmarried
21	40.0	Private	19352...	Doctorate	16.0	Married-civ-s...	Prof-speci...	Husband
22	54.0	Private	30214...	HS-grad	9.0	Separated	Other-serv...	Unmarried
23	35.0	Federal-g...	76845...	9th	5.0	Married-civ-s...	Farming-fi...	Husband
24	43.0	Private	11703...	11th	7.0	Married-civ-s...	Transport-...	Husband
25	59.0	Private	10901...	HS-grad	9.0	Divorced	Tech-sup...	Unmarried
26	56.0	Local-gov	21685...	Bachelors	13.0	Married-civ-s...	Tech-sup...	Husband
27	19.0	Private	16829...	HS-grad	9.0	Never-married	Craft-repair	Own-child
28	54.0	Private	18021...	Some-co...	10.0	Married-civ-s...	Prof-speci...	Husband
29	39.0	Private	36726...	HS-grad	9.0	Divorced	Exec-man...	Not-in-family
30	49.0	Private	19336...	HS-grad	9.0	Married-civ-s...	Craft-repair	Husband
31	23.0	Local-gov	19070...	Assoc-ac...	12.0	Never-married	Protective-...	Not-in-family

Gambar 6. setelah pre-processing

- Outlier data

Data yang outlier memiliki makna yaitu data yang memiliki nilai sangat berbeda jauh dengan nilai yang ada dalam 1 atribut tertentu. Dalam data ini permasalahan outlier data ditemukan pada atribut “Capital Gain” dan “Capital Loss”, sehingga dalam mengatasinya perlu dilakukan remove with value atau menghapus data yang menganggu tersebut, berikut penjelasan dari dilakukannya tahap outlier data :

1. Melakukan filter data dengan memilih **Filters** > **Unsupervised** > **Attribute** > **InterquartilRange.**
2. Interquartile range dilakukan untuk mendeteksi adanya data yang outlire dan yang memiliki nilai ekstrem.
3. Pilih **Filters** > **Unsupervised** > **Instance** > **RemovewithValues.** Data yang outlier akan otomatis terhapus pada tahapan ini, berikut perbandingan data sebelum dan sesudah :

l: Capital Gain Numeric	12: Capital Loss Numeric
2174.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
14084.0	0.0
5178.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0

Gambar 7.sebelum pre-processing

: Capital Gain Numeric	12: Capital Loss Numeric
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0
0.0	0.0

Gambar 8. setelah pre-processing

- Label Encoding
  Data yang akan diproses harus dirubah kedalam bentuk numeric. Salahsatunya seperti nilai atribut “Sex” yang bertype string kemudian dilakukan proses encoding dengan nilai dari “Maale” = 1, dan “Female” =0. Dalam pre processing tahap ini, dilakukan encoding label dengan menggunakan algoritma python dan *LabelEncoder*. Berikut tampilan atau hasil output dari lebel encoding :

1. Algoritma label encoding

```
labelencoder = LabelEncoder()
#data['Age_trans'] = labelencoder.fit_transform(data['Age'])
data['Workclass_trans'] = labelencoder.fit_transform(data['Workclass'])
data['Education_trans'] = labelencoder.fit_transform(data['Education'])
data['MatrialStatus_trans'] = labelencoder.fit_transform(data['MatrialStatus'])
data['Occupation_trans'] = labelencoder.fit_transform(data['Occupation'])
data['Relationship_trans'] = labelencoder.fit_transform(data['Relationship'])
data['Race_trans'] = labelencoder.fit_transform(data['Race'])
data['Sex_trans'] = labelencoder.fit_transform(data['Sex'])
data['Country_trans'] = labelencoder.fit_transform(data['Country'])
data.head()
```

Gambar 9.label encoding pada dataset Census Income

2. Hasil dari label encoding
   Hasil dari label encoding kemudian akan disimpan ke atribut baru dengan format nama “Nama atribut_trans”.

Education_trans	MatrialStatus_trans	Occupation_trans	Relationship_trans	Race_trans	Sex_trans
11	0	5	1	4	1
1	2	5	0	2	1
9	2	9	5	2	0
12	2	3	5	4	0
11	2	3	0	4	1

Gambar 10. hasil encoding dataset Census Income

- Discretization
  Discretization digunakan untuk memberikan rentan nilai pada isi dari data tesebut sehingga nantinya akan mudah untuk dipahami dan digunakan untuk mengurangi noise pada data karena data-data tersebur sebagian memiliki angka yang besar. Dalam proses ini atribut data yang didiscretization adalah “Age” dan “Fnlwgt”.
  1. Algoritma Discrerization pada atribut “Age”
     Dalam tahap ini pengelompokan data akan terbagi menjadi 5 bagian dengan rentan nilai yang sudah ditentukan.

```
enc = KBinsDiscretizer(n_bins=5, encode='ordinal', strategy='uniform')
X_binned = enc.fit_transform(data[['Age']])
data['trans_age'] = X_binned
data.head()
```

Gambar 11. Discretization pada atribut "Age"

2. Hasil Discrerization

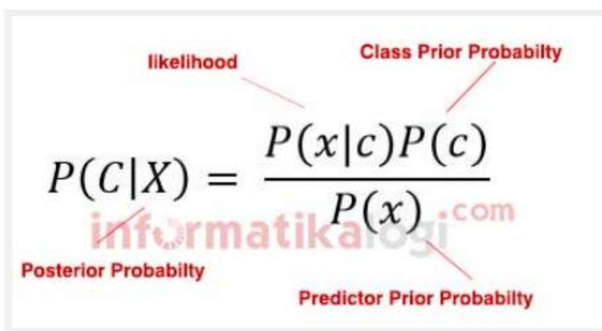
Age
38
53
28
37
52

Gambar 12. sebelum Discretization

trans_age
1.0
2.0
0.0
1.0
2.0

Gambar 13. sesudah Discretization

3. Klasifikasi
   Algoritma yang digunakan untuk pengklasifikasian data-data tersebut adalah Naïve Bayes. Naïve bayes merupakan metode klasifikasi dengan menggunakan konsep probabilitas dan statistika. Algoritma Naïve Bayes memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya, sehingga dikenal dengan “teorema bayes”. Keuntungan dari penggunaan algoritma ini adalah hanya dibutuhkan jumlah data pelatihan(trining) yang kecil untuk menentukan estimasi parameter yang dibutuhkan dalam proses pengklasifikasian. Berikut penjelasan dari rumus Persamaan Teorema Bayes :



Gambar 14. rumus Naive Bayes

Keterangan :

- X : Data class yang belum diketahui.
- C : Hipotesis data menupakan suatu class spesifik.
- P(c|x) : probabilitas hipotesis berdasarkan kondisi(posterion probability)

- P(c) : probabilitas hipotesis
- P(x|c) : probabilitas berdasarkan kondisi pada hipotesis
- P(x) : probabilitas c

3.1 Breast Cancer

Klasifikasi yang dilakukan pada dataset Breast Cancer menggunakan bahasa pemrograman python. Dengan data input sebanyak 9 atribut dan record sebanyak 285, dan data target sebanyak 1 atribut dengan record 285. Dalam proses ini telah ditentukan data training sebanyak 80% dari jumlah total yaitu sebanyak 228, dan data testing yang diambil hanya 20% dari jumlah total yaitu sebanyak 58. kemudian hasil akurasi yang didapat dari klasifikasi dengan algoritma Naïve Bayes adalah 0,79, jika dala bentuk persen menjadi 79%.

Berikut beberapa detai penjelasan dari proses klasifikasi data Breast Cancer :

- Setelah dilakukan pre-processing, data dapat langsung digunakan, dengan menentukan data input dan data target. Berikut pemilihan datanya :

```
X = data[['Class_trns','Age_trans','Menopause_trans','TumorSize_trans','InvNodes_trans','NodeCaps_trans
y = data[['Irradiant_trans']]
```

Gambar 15. Pemilihan data train dan data test

- Setelah dipilih dilakukan proses pembagian data training dan data testing. Seeprti yang sudah dijelaskan diatas, data training yag digunakan sebanyak 80% dan data testing sebnayak 20%. Berikut algoritmanya :

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=1)
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
```

```
(228, 9)
(228, 1)
(58, 9)
(58, 1)
```

Gambar 16. pembagian data train dan data test

- Data yang sudah dibagi selanjutnya dapat dilakukan proses klasifikasi dengan penjelasan algoritma sebagai berikut:

```
NaiveBayes = MultinomialNB().fit(X_train,np.ravel(y_train,order='c'))
print(NaiveBayes)
```

Gambar 17. Library Naive Bayes

- Data yang sudah diklasifikasikan kemudian akan memperoleh hasil dan tingkat akurasi sebagai beikut :

```
prediction = NaiveBayes.predict(X_test)
print(prediction)

from sklearn.metrics import classification_report
print(classification_report(y_test, prediction))
```

```
[0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0
 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 1 1 0]
              precision    recall  f1-score   support

           0       0.83      0.90      0.86        42
           1       0.67      0.50      0.57        16

    accuracy                           0.79        58
   macro avg       0.75      0.70      0.72        58
weighted avg       0.78      0.79      0.78        58
```

Gambar 18. hasil prediksi dan akurasi

3.2 Census Income

Klasifikasi yang dilakukan pada dataset Census Income menggunakan bahasa pemrograman python. Dengan data input sebanyak 12 atribut dan record sebanyak 285, dan data target sebanyak 1 atribut dengan record 32.561. Dalam proses ini telah ditentukan data training sebanyak 80% dari jumlah total yaitu sebanyak19440, dan data testing yang diambil hanya 20% dari jumlah total yaitu sebanyak 4861. kemudian hasil akurasi yang didapat dari klasifikasi dengan algoritma Naïve Bayes adalah 0,81, jika dala bentuk persen menjadi 81%. Berikut beberapa detai penjelasan dari proses klasifikasi data Census Income:

- Setelah dilakukan pre-processing, data dapat langsung digunakan, dengan menentukan data input dan data target. Berikut pemilihan datanya :

```
X = data[['trans_age','Workclass_trans','trans_Fnlwgt','Education_trans','Education-Num','MatrialStatus
y = data[['Target']]
```

Gambar 19. Pemilihan data train dan data test

- Setelah dipilih dilakukan proses pembagian data training dan data testing. Seeprti yang sudah dijelaskan diatas, data training yag digunakan sebanyak 80% dan data testing sebnayak 20%. Berikut algoritmanya :

```
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2, random_state=1)
print(X_train.shape)
print(y_train.shape)
print(X_test.shape)
print(y_test.shape)
```

```
(19440, 12)
(19440, 1)
(4861, 12)
(4861, 1)
```

Gambar 20. pembagian data train dan data test

- Data yang sudah dibagi selanjutnya dapat dilakukan proses klasifikasi dengan penjelasan algoritma sebagai berikut:

```
NaiveBayes = MultinomialNB().fit(X_train,np.ravel(y_train,order='c'))
print(NaiveBayes)
```

Gambar 21. Library Naive Bayes

- Data yang sudah diklasifikasikan kemudian akan memperoleh hasil dan tingkat akurasi sebagai beikut :

```
prediction = NaiveBayes.predict(X_test)
print(prediction)

from sklearn.metrics import classification_report
print(classification_report(y_test, prediction))
```

```
['<=50K' '<=50K' '<=50K' ... '>50K' '<=50K' '<=50K']
              precision    recall  f1-score   support

       <=50K       0.87      0.90      0.88     15449
        >50K       0.54      0.48      0.51      3991

    accuracy                           0.81     19440
   macro avg       0.71      0.69      0.70     19440
weighted avg       0.80      0.81      0.81     19440
```

Gambar 22. hasil prediksi dan akurasi