Laporan Akhir Data Mining

Analisis Perbandingan Metode Klasifikasi pada Dataset Kanker Payudara Menggunakan Feature Selection Principal Component Analysis (PCA)

Disusun Oleh :

Muhammad Fikri Septiawan 1301160789

Program Studi S1 Informatika

Fakultas Informatika

Telkom University

2019

DAFTAR ISI

1. Deskripsi Data

Dataset yang digunakan adalah dataset eksperimen kanker payudara Wisconsin (Diagnostik) dan dapat dilihat maupun diunduh pada website *UCI machine learning*. Dataset tersebut memiliki 699 baris, fitur sebanyak 10 dan *class* seperti pada gambar 1.

1	id number	Clump Thicknes	Uniformity of Cell Size	Uniformity of Cell Shape	Marginal Adhesion	Single Epithelial Cell Size	Bare Nuclei	Bland Chromatin	Normal Nucleoli	Mitoses	Class
2	1000025	5	1	1	1	2	1	3	1	1	2
3	1002945	5	4	4	5	7	10	3	2	1	2
4	1015425	3	1	1	1	2	2	3	1	1	2
5	1016277	6	8	8	1	3	4	3	7	1	2
6	1017023	4	1	1	3	2	1	3	1	1	2
7	1017122	8	10	10	8	7	10	9	7	1	4
8	1018099	1	1	1	1	2	10	3	1	1	2
9	1018561	2	1	2	1	2	1	3	1	1	2
10	1033078	2	1	1	1	2	1	1	1	5	2

	A	B	C	D	E	F	G	H	I	J	K
681	1368882	2	1	1	1	2	1	1	1	1	2
682	1369821	10	10	10	10	5	10	10	10	7	4
683	1371026	5	10	10	10	4	10	5	6	3	4
684	1371920	5	1	1	1	2	1	3	2	1	2
685	466906	1	1	1	1	2	1	1	1	1	2
686	466906	1	1	1	1	2	1	1	1	1	2
687	534555	1	1	1	1	2	1	1	1	1	2
688	536708	1	1	1	1	2	1	1	1	1	2
689	566346	3	1	1	1	2	1	2	3	1	2
690	603148	4	1	1	1	2	1	1	1	1	2
691	654546	1	1	1	1	2	1	1	1	8	2
692	654546	1	1	1	3	2	1	1	1	1	2
693	695091	5	10	10	5	4	5	4	4	1	4
694	714039	3	1	1	1	2	1	1	1	1	2
695	763235	3	1	1	1	2	1	2	1	2	2
696	776715	3	1	1	1	3	2	1	1	1	2
697	841769	2	1	1	1	2	1	1	1	1	2
698	888820	5	10	10	3	7	3	8	10	2	4
699	897471	4	8	6	4	3	4	10	6	1	4
700	897471	4	8	8	5	4	5	10	4	1	4
701											

Gambar 1. Dataset

Penjelasan tiap fitur dan *class :*

- *Id number* : Yaitu id saat dilakukan eksperimen
- *Clump Thickness* : Ketebalan gumpalan (Rentang nilai 1-10)
- *Uniformity of Cell Size* : Keseragaman Ukuran Sel (Rentang nilai 1-10)
- *Uniformity of Cell Shape* : Keseragaman Bentuk Sel (Rentang nilai 1-10)
- *Marginal Adhesion* : Adhesi Marginal (Rentang nilai 1-10)
- *Single Epithelial Cell Size* : Ukuran Sel Epitel Tunggal (Rentang nilai 1-10)
- *Bare Nuclei* : *Bare Nuclei* (Rentang nilai 1-10)

- *Bland Chromatin* : Chromatin lembut (Rentang nilai 1-10)
- *Normal Nucleoli* : Nukleoli normal (Rentang nilai 1-10)
- *Mitoses* : Jenis pembelahan sel yang menghasilkan dua sel anak masing-masing memiliki jumlah dan jenis kromosom yang sama dengan inti induk, khas dari pertumbuhan jaringan biasa (Rentang nilai 1-10)
- *Class* : Kelas atau label, terdapat 2 kelas yaitu 2 (benign atau jinak) dan 4 (malignant atau ganas)

2. Praproses Data

- **Pemilihan Baris Data**

  Data yang digunakan dalam percobaan ini ada sebanyak 699 baris data (semuanya digunakan) dan melakukan convert untuk class 2 menjadi benign (jinak) dan class 4 menjadi malignant (ganas).

- **Split Dataset**

  Adapun dalam percobaan ini menggunakan 80% data *train* dan 20% data test

- **Pemilihan Fitur**

  Proses ini dilakukan untuk memilih fitur apa saja yang akan dipilih untuk keperluan klasifikasi. Dimana Fitur yang dipilih disini yaitu fitur *Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses*. Sementara itu fitur id number didrop karena dianggap tidak penting dalam klasifikasi. Selain itu, pemilihan fitur pada percobaan ini yaitu menggunakan PCA (*Principal Component Analysis*). Penulis melakukan percobaan PCA dengan 2 komponen dan 3 komponen.

- **Cleaning**

  Proses cleaning dataset diperlukan untuk mengatasi missing values. Setiap missing values pada dataset akan diisi dengan nilai modus (data yang sering muncul). Pada percobaan disini terdapat missing value pada fitur *Bare Nuclei* yang diisi dengan ?(tanda tanya) dan dapat dilihat pada gambar 2.

25	1057013	8	4	5	1	2 ?	7	3	1	4
42	1096800	6	6	6	9	6 ?	7	8	1	2
141	1183246	1	1	1	1	1 ?	2	1	1	2
147	1184840	1	1	3	1	2 ?	2	1	1	2
160	1193683	1	1	2	1	3 ?	1	1	1	2

Gambar 2. Missing value

3. Hasil Percobaan

- **Principal Component Analysis (PCA)**

Out[58]:

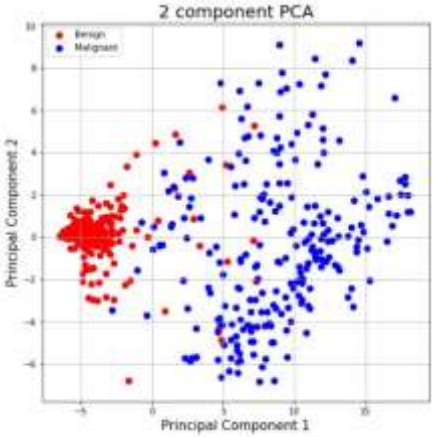
	PC1	PC2	Class
0	-4.418342	0.032549	Benign
1	4.861916	-4.869842	Benign
2	-4.575277	-0.629810	Benign
3	5.165415	3.406013	Benign
4	-4.053561	-0.105016	Benign
5	15.067319	-0.528628	Malignant
6	-1.554196	-6.795075	Benign
7	-4.921075	0.383009	Benign
8	-5.402775	0.837241	Benign
9	-4.804405	0.308262	Benign
10	-5.361556	0.096435	Benign

Gambar 3. PCA 2 Komponen

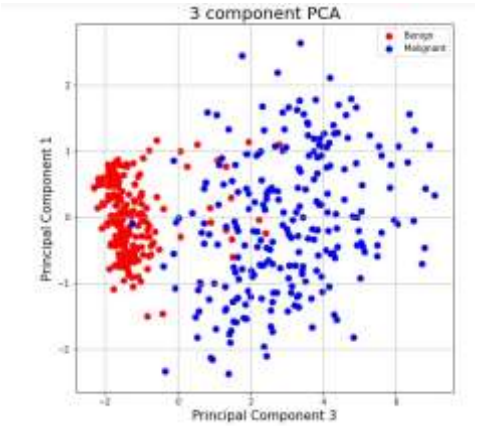
Out[74]:

	pc1	pc2	pc3	Class
0	-1.456220	-0.110210	-0.574438	Benign
1	1.466279	-0.544894	0.283038	Benign
2	-1.579311	-0.074854	0.037413	Benign
3	1.505247	-0.568853	-0.612984	Benign
4	-1.330551	-0.089657	0.027402	Benign
5	5.054140	-1.542614	0.476466	Malignant
6	-1.057400	-0.518582	0.642164	Benign
7	-1.651934	0.016774	0.333214	Benign
8	-1.526659	2.354349	-0.037983	Benign
9	-1.580810	0.019803	-0.342779	Benign
10	-2.038715	0.023250	0.601532	Benign

Gambar 4. PCA 3 Komponen

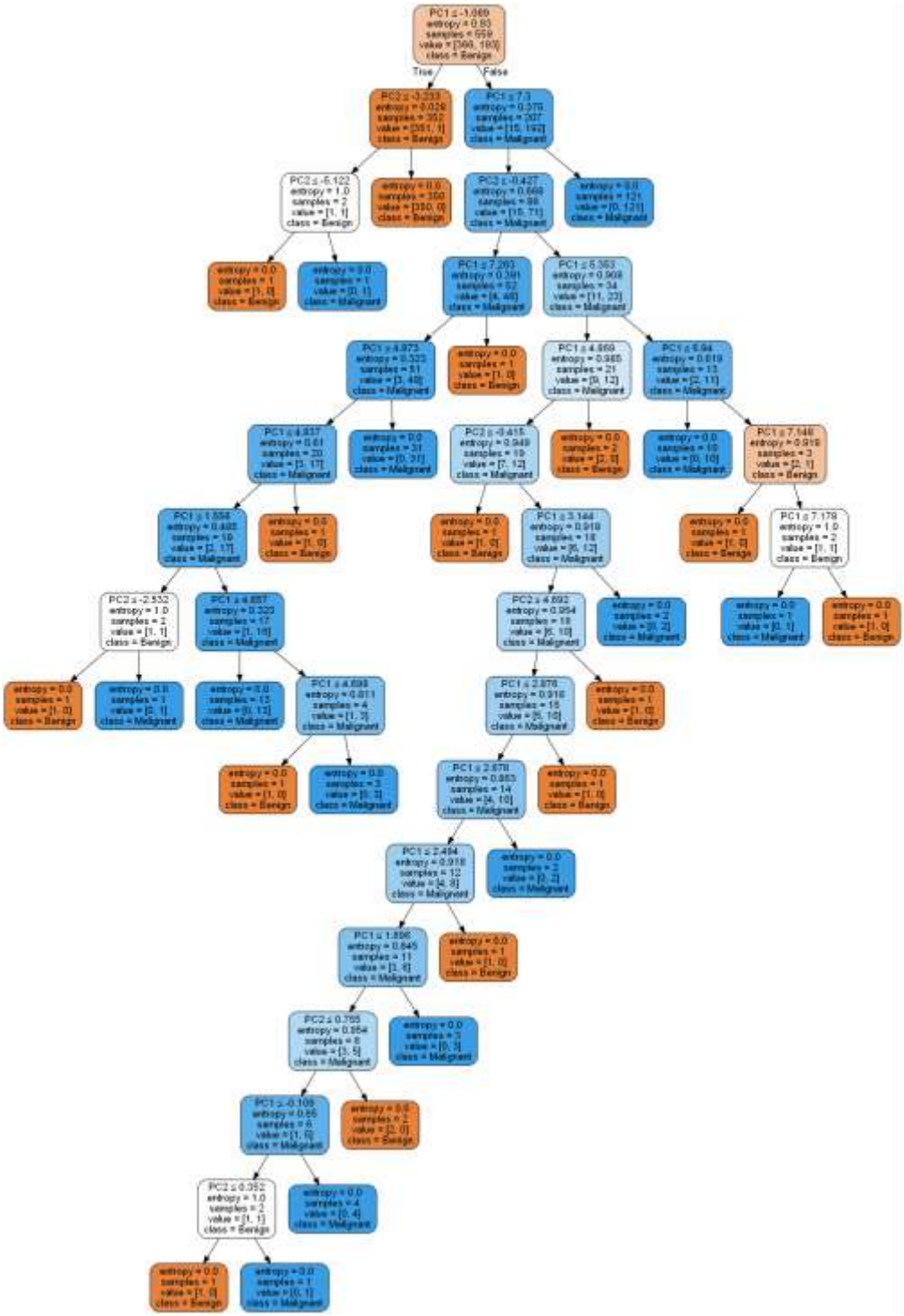


Gambar 5. Plot PCA 2 Komponen

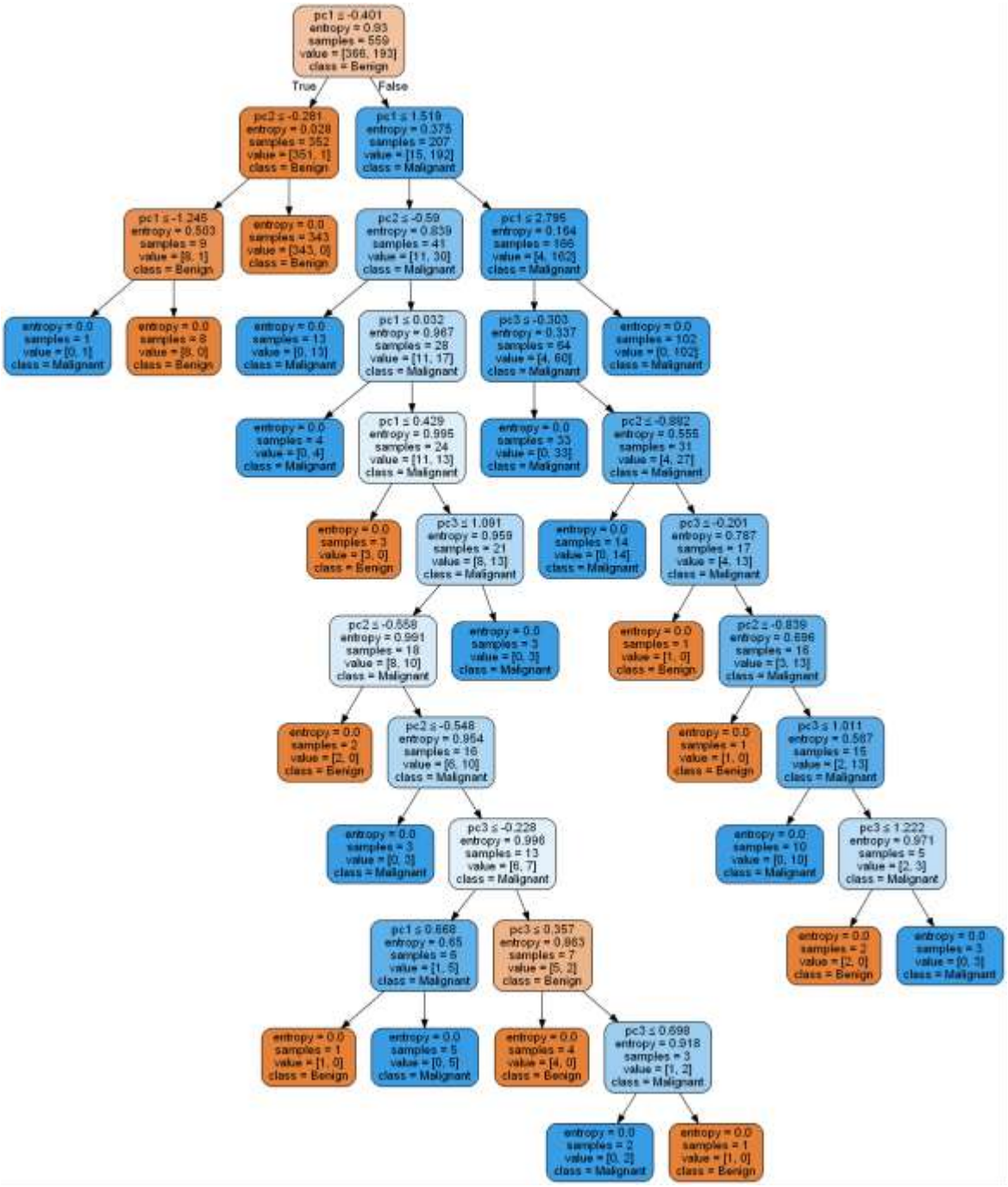


Gambar 6. Plot PCA

- **Metode Klasifikasi**
  - **Algoritma ID3**

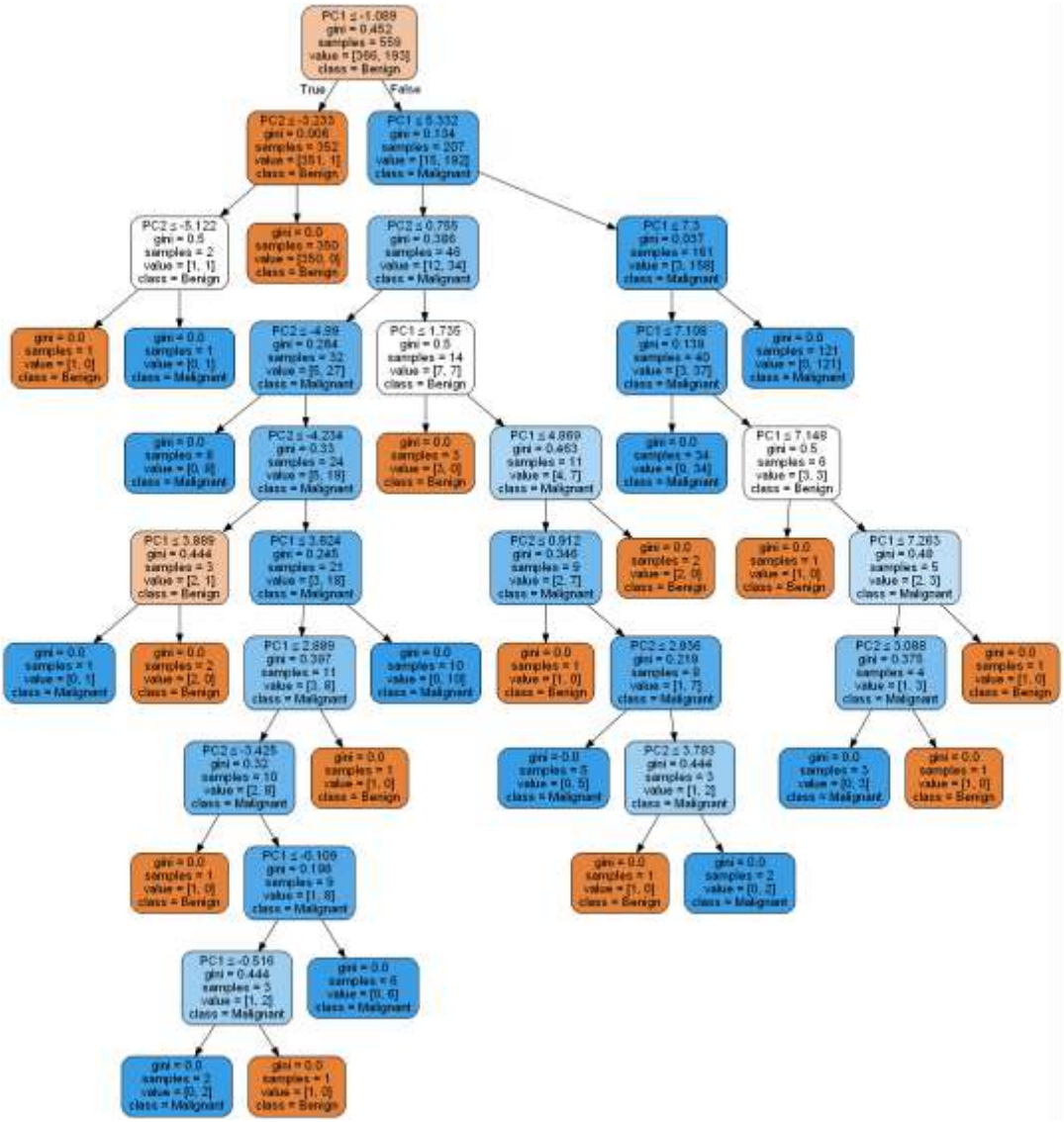


Gambar 3. Tree Algoritma ID3 PCA 2 Komponen

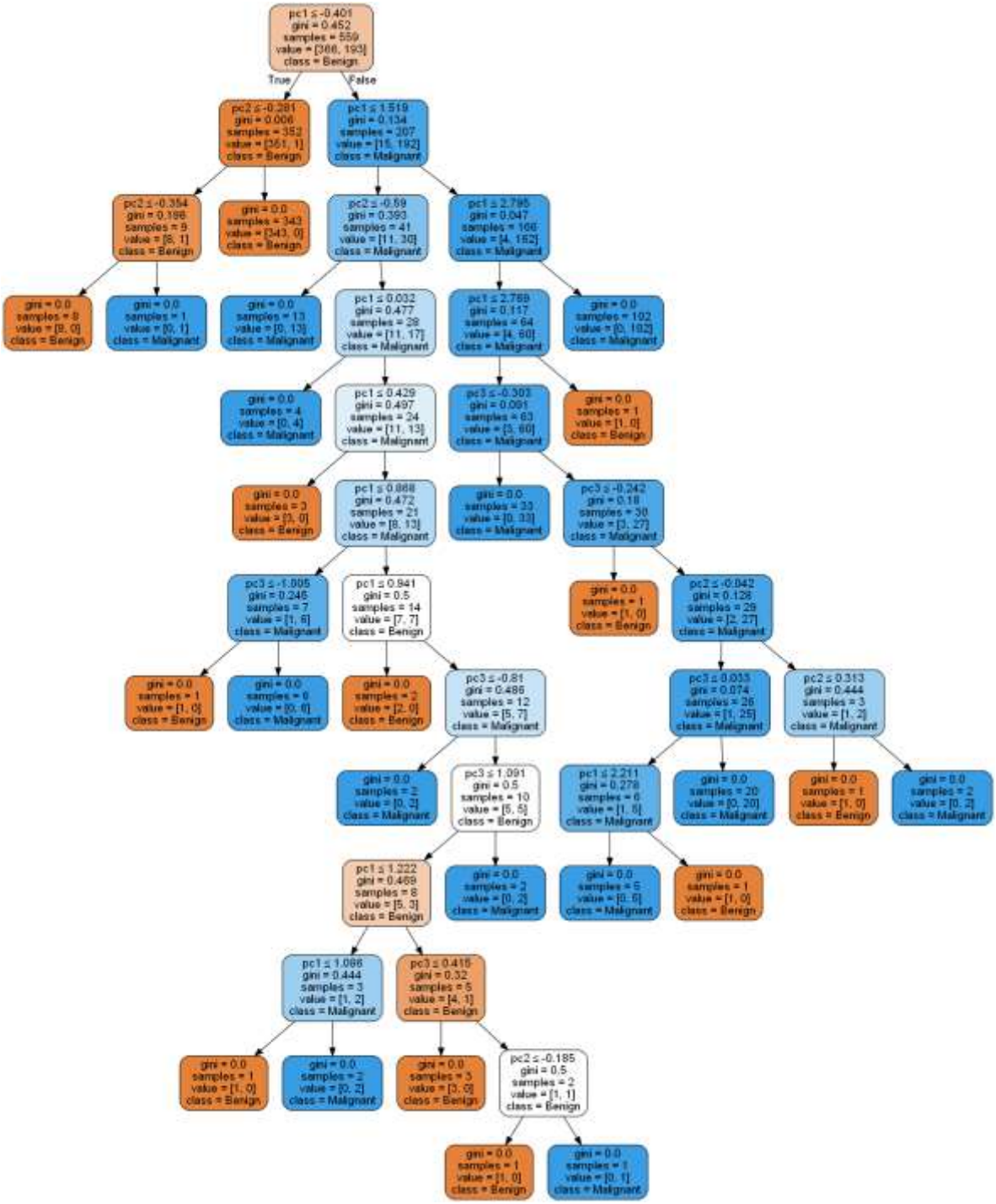


Gambar 8. Tree Algoritma ID3 PCA 3 Komponen

- **Algoritma Cart**



Gambar 9. Tree Algoritma Cart PCA 2 Komponen



Gambar 10. Tree Algoritma Cart PCA 3 Komponen

- Algoritma Naive Bayes (Gaussian Naive Bayes)

```
In [74]: y_pred

Out[74]: array(['Malignant', 'Benign', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Benign', 'Malignant', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Malignant', 'Malignant',
                'Benign', 'Malignant', 'Malignant', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Malignant', 'Benign', 'Malignant', 'Malignant', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Malignant', 'Benign', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Malignant', 'Malignant',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Benign', 'Malignant',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Malignant',
                'Malignant', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Malignant',
                'Benign', 'Benign', 'Benign', 'Benign', 'Malignant', 'Malignant',
                'Benign', 'Malignant', 'Malignant', 'Malignant'], dtype='<U9')
```

Gambar 11. Hasil Prediksi Algoritma Gaussian Naive Bayes PCA 2 Komponen

```
In [45]: y_pred

Out[45]: array(['Malignant', 'Benign', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Benign', 'Malignant', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Malignant', 'Malignant',
                'Benign', 'Malignant', 'Malignant', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Malignant', 'Benign',
                'Malignant', 'Benign', 'Malignant', 'Malignant', 'Benign',
                'Malignant', 'Malignant', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Malignant', 'Malignant', 'Benign', 'Benign',
                'Malignant', 'Benign', 'Benign', 'Malignant', 'Malignant',
                'Malignant', 'Benign', 'Benign', 'Benign', 'Benign', 'Malignant',
                'Malignant', 'Malignant', 'Benign', 'Benign', 'Benign', 'Benign',
                'Benign', 'Benign', 'Benign', 'Benign', 'Benign', 'Malignant',
                'Benign', 'Benign', 'Benign', 'Benign', 'Malignant', 'Malignant',
                'Benign', 'Benign', 'Malignant', 'Benign', 'Malignant', 'Benign',
                'Benign', 'Benign', 'Benign', 'Malignant', 'Benign', 'Benign',
                'Benign', 'Benign', 'Malignant', 'Malignant', 'Benign',
                'Malignant', 'Malignant', 'Malignant'], dtype='<U9')
```

Gambar 12. Hasil Prediksi Algoritma Naive Bayes PCA 3 Komponen

3.1 Evaluasi Performa Metode Klasifikasi

- **Confusion Matrix**

```
[[91  1]
 [ 8 40]]
              precision    recall  f1-score   support

      Benign       0.92      0.99      0.95        92
   Malignant       0.98      0.83      0.90        48

   micro avg       0.94      0.94      0.94       140
   macro avg       0.95      0.91      0.93       140
weighted avg       0.94      0.94      0.93       140
```

Gambar 13. Confusion Matrix Cart PCA 2 Komponen

```
[[90  2]
 [ 4 44]]
              precision    recall  f1-score   support

      Benign       0.96      0.98      0.97        92
   Malignant       0.96      0.92      0.94        48

   micro avg       0.96      0.96      0.96       140
   macro avg       0.96      0.95      0.95       140
weighted avg       0.96      0.96      0.96       140
```

Gambar 14. Confusion Matrix Cart PCA 3 Komponen

```
[[91  1]
 [ 7 41]]
              precision    recall  f1-score   support

      Benign       0.93      0.99      0.96        92
   Malignant       0.98      0.85      0.91        48

   micro avg       0.94      0.94      0.94       140
   macro avg       0.95      0.92      0.93       140
weighted avg       0.94      0.94      0.94       140
```

Gambar 15. Confusion Matrix ID3 PCA 2 Komponen

```
[[90  2]
 [ 0 48]]
              precision    recall  f1-score   support

      Benign       1.00      0.98      0.99        92
   Malignant       0.96      1.00      0.98        48

   micro avg       0.99      0.99      0.99       140
   macro avg       0.98      0.99      0.98       140
weighted avg       0.99      0.99      0.99       140
```

Gambar16. Confusion Matrix ID3 PCA 3 Komponen

```
              precision    recall  f1-score   support

      Benign       0.98      0.96      0.97        96
   Malignant       0.91      0.95      0.93        44

   micro avg       0.96      0.96      0.96       140
   macro avg       0.95      0.96      0.95       140
weighted avg       0.96      0.96      0.96       140

[[92  4]
 [ 2 42]]
```

Gambar 17. Confusion Matrix Naive Bayes PCA 2 Komponen

```
              precision    recall  f1-score   support

      Benign       0.98      0.96      0.97        96
   Malignant       0.91      0.95      0.93        44

   micro avg       0.96      0.96      0.96       140
   macro avg       0.95      0.96      0.95       140
weighted avg       0.96      0.96      0.96       140

[[92  4]
 [ 2 42]]
```

Gambar 18. Confusion Matrix ID3 PCA 3 Komponen

- **Akurasi**

```
In [108]: print("Akurasi:",metrics.accuracy_score(y_test, y_pred))
          Akurasi: 0.9357142857142857
```

Gambar 19. Akurasi Cart PCA 2 Komponen

```
In [47]: # Model Accuracy, how often is the classifier correct?
         print("Akurasi:",metrics.accuracy_score(y_test, y_pred))
         Akurasi: 0.9571428571428572
```

Gambar 20. Akurasi Cart PCA 3 Komponen

```
In [114]: print("Akurasi:",metrics.accuracy_score(y_test, y_pred))
          Akurasi: 0.9428571428571428
```

Gambar 21. Akurasi ID3 PCA 2 Komponen

```
In [36]: # Model Accuracy, how often is the classifier correct?
         print("Akuras:",metrics.accuracy_score(y_test, y_pred))
         Akuras: 0.9857142857142858
```

Gambar 22. Akurasi ID3 PCA 3 Komponen

```
In [125]: from sklearn import metrics

          print("Akurasi:",metrics.accuracy_score(y_test, y_pred))
          Akurasi: 0.9571428571428572
```

Gambar 23. Akurasi Naive Bayes PCA 2 Komponen

```
In [92]: from sklearn import metrics

         print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
         Accuracy: 0.9571428571428572
```

Gambar 24. Akurasi Naive Bayes PCA 3 Komponen

- ***Comparative Analysis***
  - **PCA 2 Komponen**

	CCI	*ICI*	*RMS*
Cart	*93,5714%*	*6,4286%*	*0,64*
ID3	*94,2857%*	*5,7143%*	*0,57*
Naïve Bayes	*95,7143%*	*4,2857%*	*0,42*

Tabel 1. Comparative Analysis PCA 2 Komponen

  - **PCA 3 Komponen**

	CCI	*ICI*	*RMS*
Cart	*95,7143%*	*4,2857%*	*0,42*
ID3	*98,7143%*	*1,2857%*	*0,12*
Naïve Bayes	*95,7143%*	*4,2857%*	*0,42*

Tabel 2. Comparative Analysis PCA 3 Komponen

4. Analisis Perbandingan Metode Klasifikasi

Didapat metode terbaik dari 3 metode klasifikasi adalah algoritma ID3 dengan PCA 3 komponen.