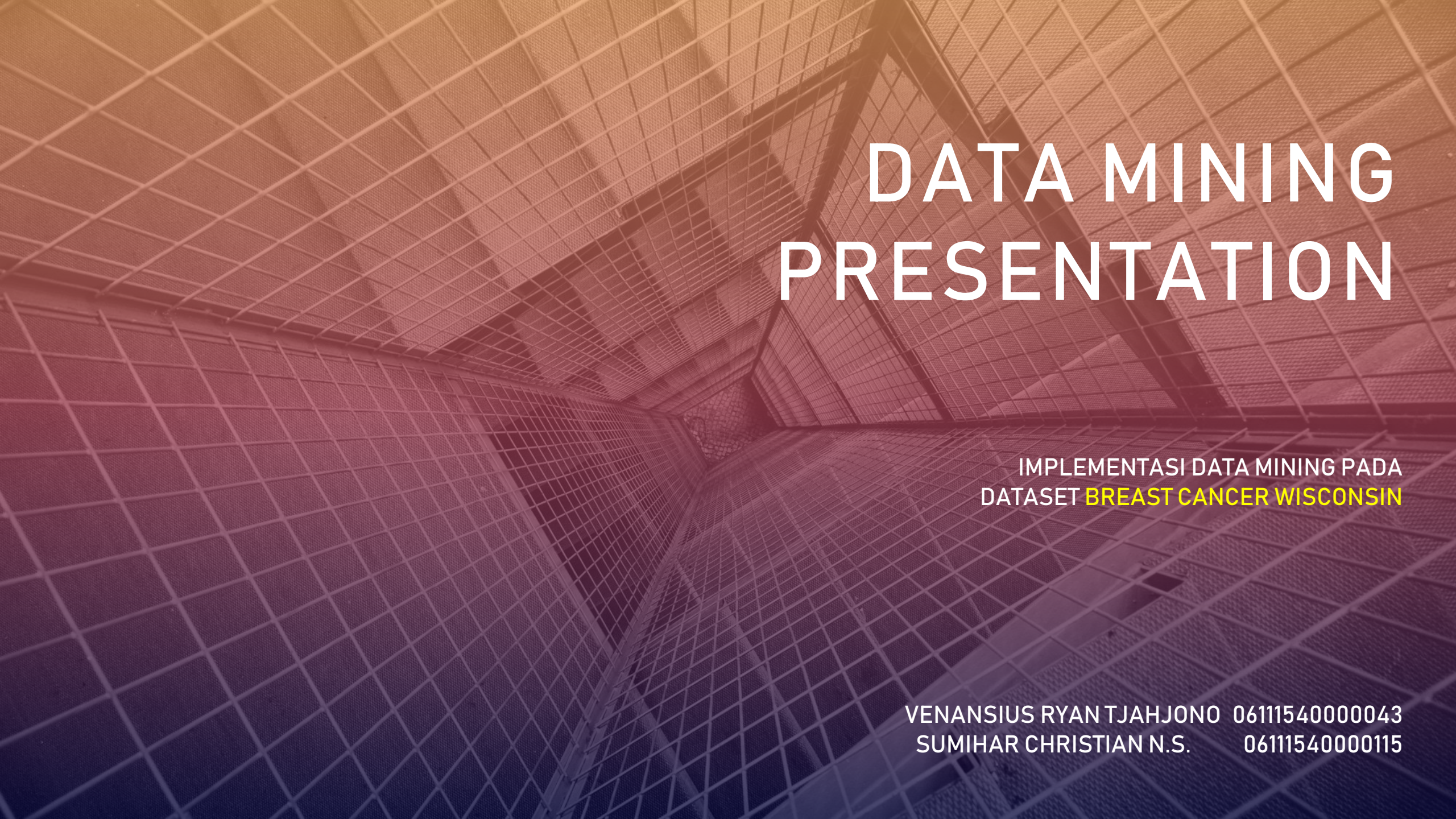
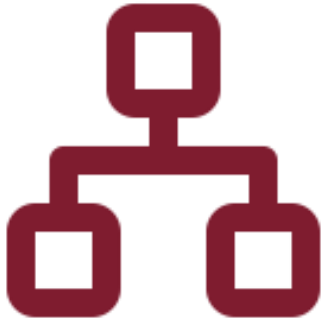


DATA MINING PRESENTATION

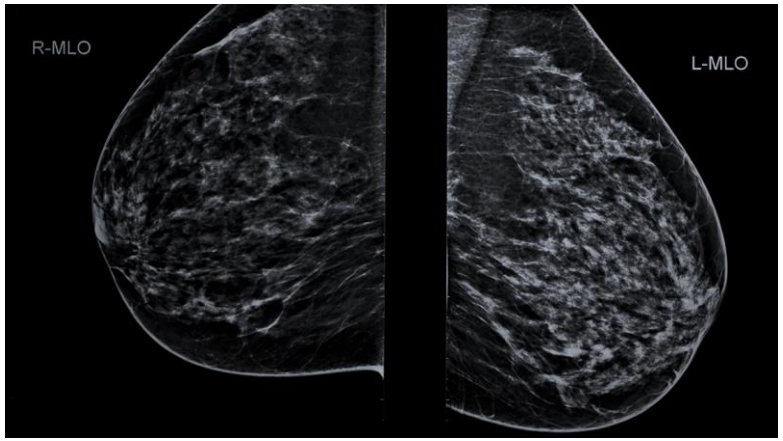
IMPLEMENTASI DATA MINING PADA
DATASET BREAST CANCER WISCONSIN

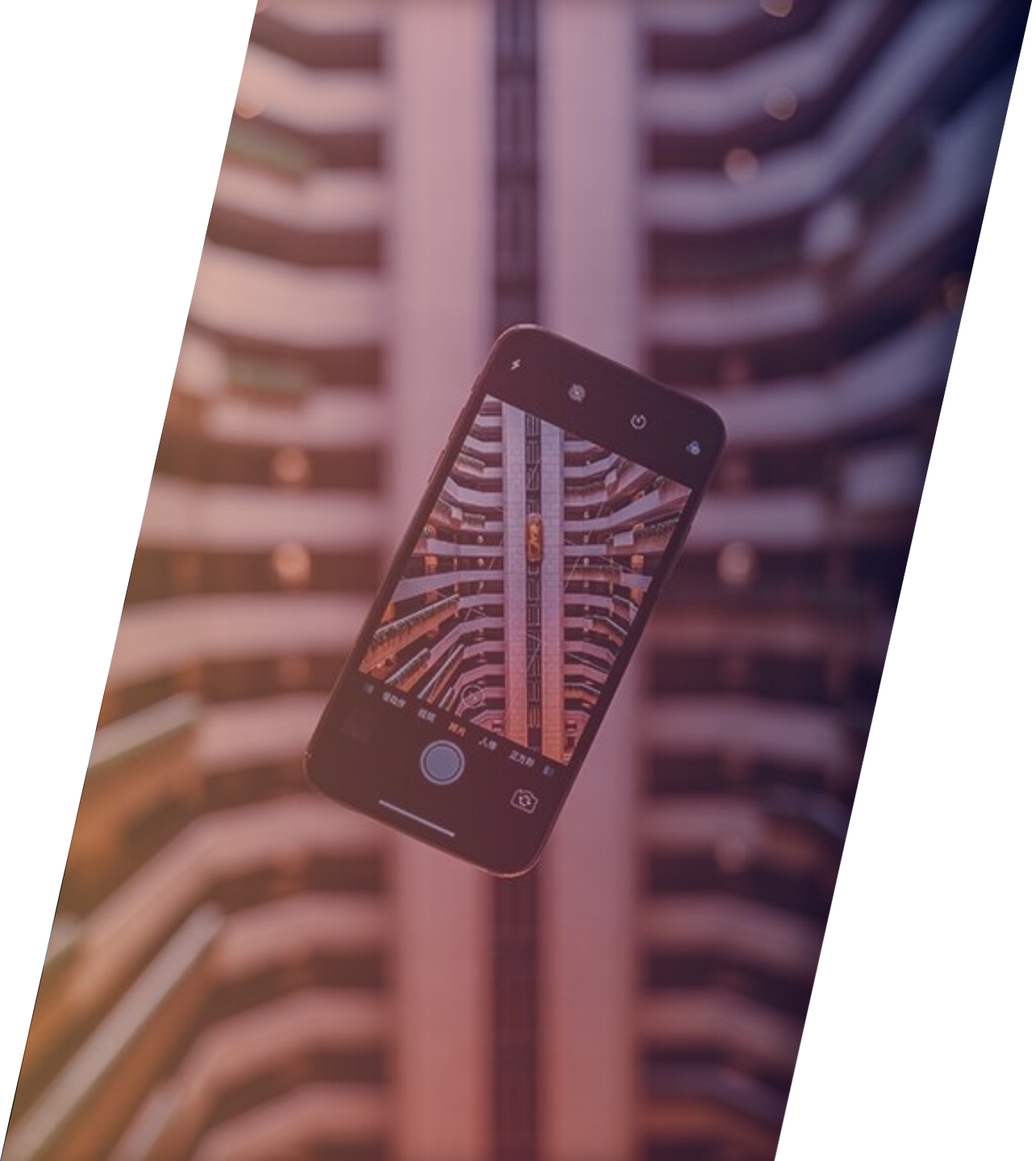
VENANSIUS RYAN TJAHJONO 06111540000043
SUMIHAR CHRISTIAN N.S. 06111540000115

LATAR BELAKANG



2





RUMUSAN MASALAH

CARA PREPROCESSING DATA

ANALISIS DATASET DENGAN TASK DATA MINING

CROSS VALIDATION DENGAN MULTI LAYER PERCEPTRON

DATASETS

https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

Name	Last modified	Size	Description
Parent Directory		-	
Index	03-Dec-1996 04:07	326	
breast-cancer-wisconsin.data	16-Jul-1992 10:15	19K	
breast-cancer-wisconsin.names	16-Jul-1992 14:13	5.5K	
unformatted-data	16-Jul-1992 06:17	21K	
wdbc.data	05-Feb-1996 11:04	121K	
wdbc.names	05-Feb-1996 11:04	4.6K	
wpbc.data	01-Feb-1996 16:00	43K	
wpbc.names	01-Feb-1996 16:00	5.5K	

Apache/2.2.15 (CentOS) Server at archive.ics.uci.edu Port 443

ATRIBUT DATA (mean, se, worst)

1. ID number
2. Diagnosis (M = malignant, B = benign)

3 - 32 Ten real-valued features are computed for each cell nucleus:

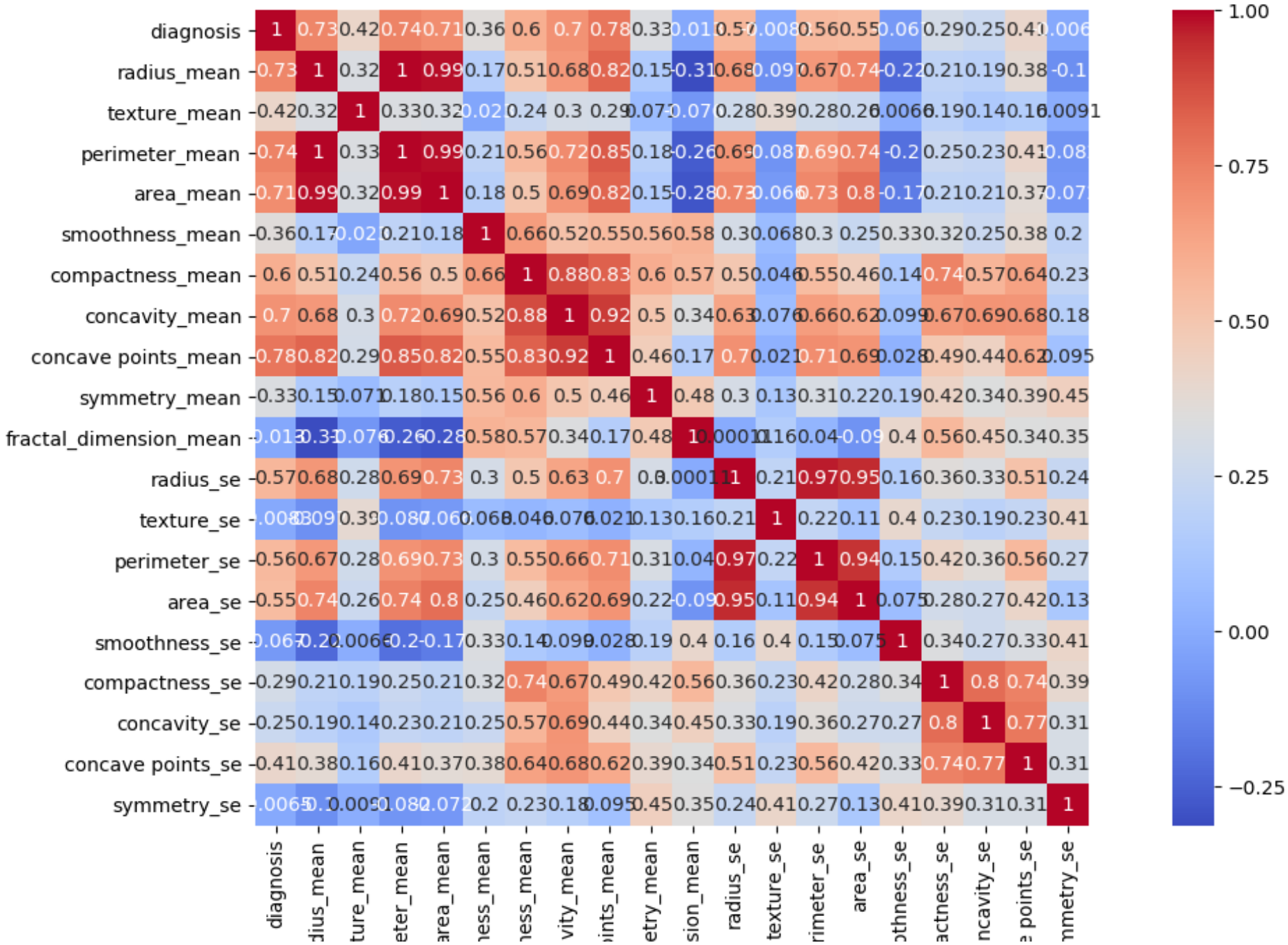
- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

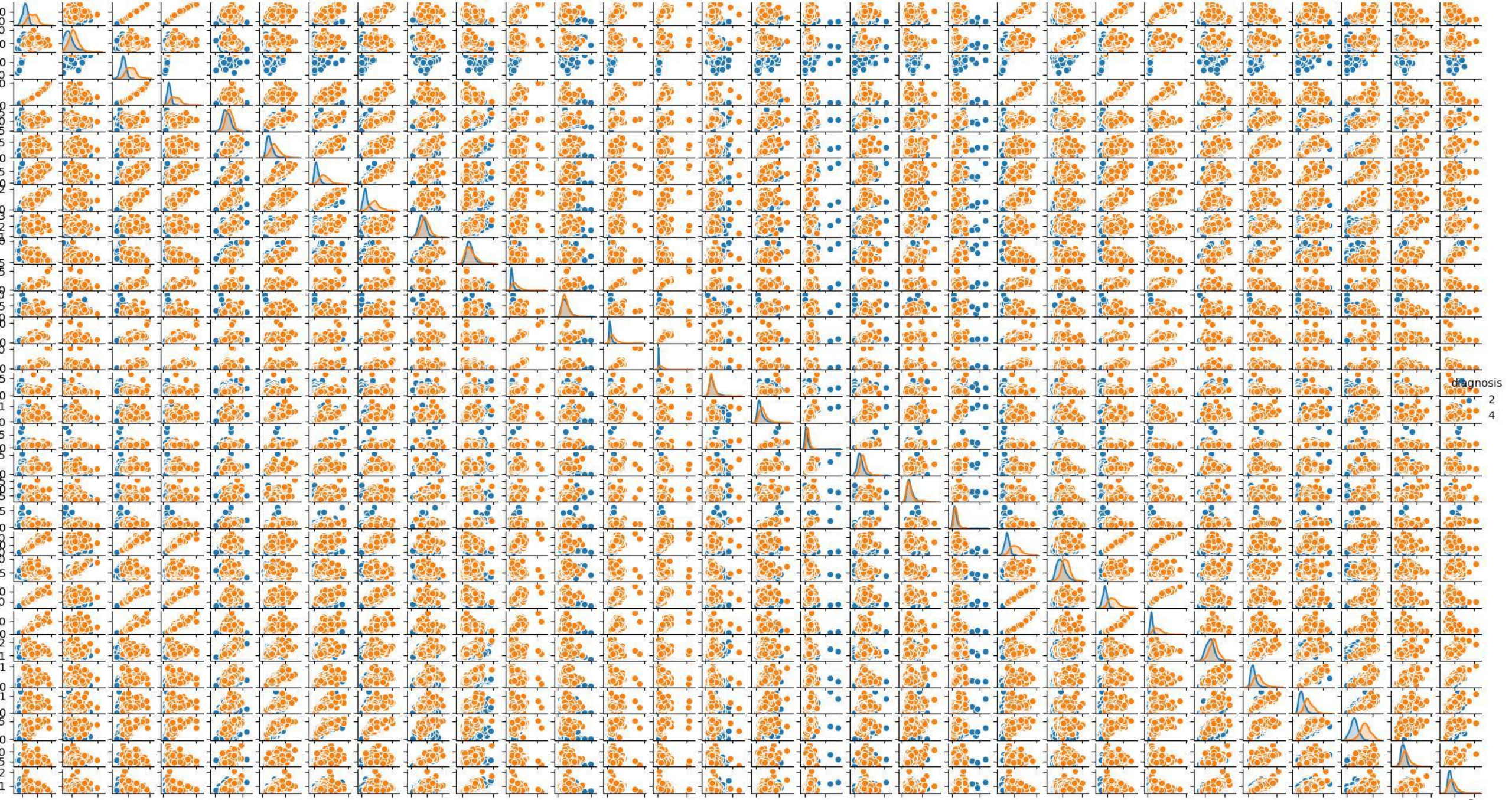
```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('data.csv', header=0)

#PREPROCESSING DATA
# dataset.replace('?', -99999, inplace=True) #-9999 biar
outlier, gak masuk ke grafik
dataset.drop("id",1)
mapping={'M':4, 'B':2}
print(dataset.shape)
dataset['diagnosis'] = dataset['diagnosis'].map(mapping)
X = dataset.iloc[:, 1:31].values # parameter yang mau di
train
y = dataset.iloc[:, 1].values # target
```

PREPROCESSING DATA

SUATU PROSES/LANGKAH YANG DILAKUKAN UNTUK MEMBUAT DATA MENTAH MENJADI DATA YANG BERKUALITAS(INPUT YANG BAIK UNTUK DATA MINING TOOLS).

PADA DATA YANG DIPILIH, DITANGANI NILAI OUTLIER SAJA SEBAB TIDAK ADA *MISSING VALUE*.


diagnosis
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave points_se
symmetry_se
1.00
0.75
0.50
0.25
0.00
−0.25


diagnosis
2
4

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('data.csv', header=0)

#PREPROCESSING DATA
# dataset.replace('?', -99999, inplace=True) #-9999 biar
outlier, gak masuk ke grafik
dataset.drop("id",1)
mapping={'M':4, 'B':2}
print(dataset.shape)
dataset['diagnosis'] = dataset['diagnosis'].map(mapping)
X = dataset.iloc[:, 1:31].values # parameter yang mau di
train
y = dataset.iloc[:, 1].values # target
```

PLOT DATASET

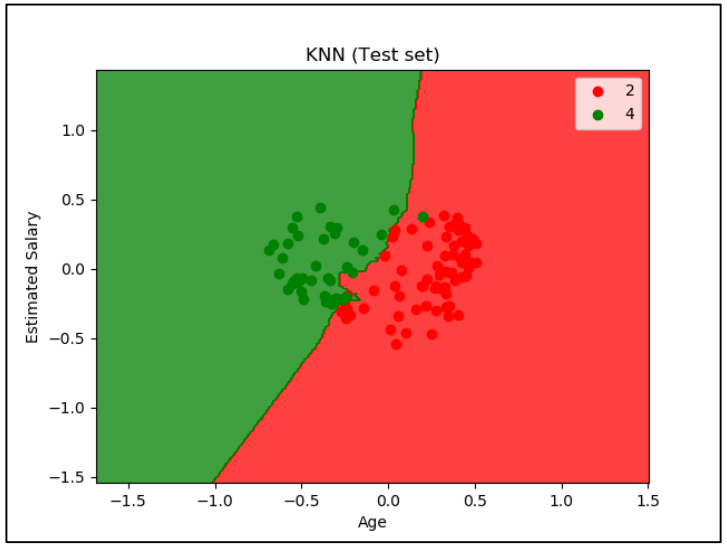
PROSES PLOTTING DILAKUKAN DALAM BENTUK EXPLANATORY DATA ANALYSIS.

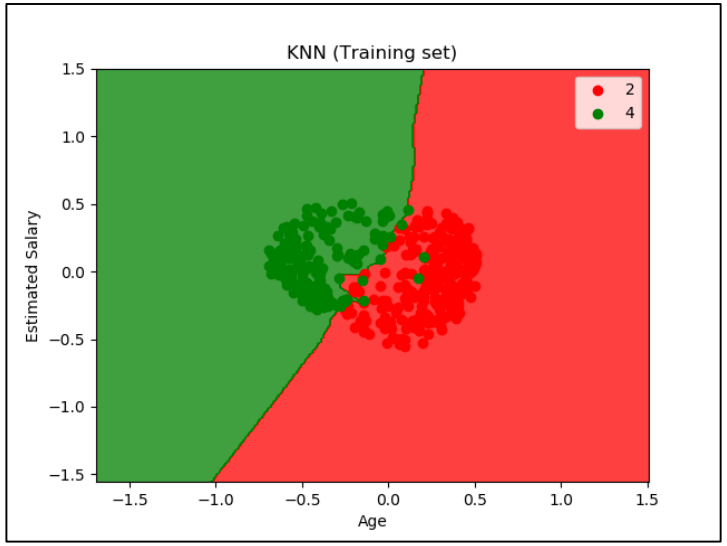
CLASSIFICATION

PROSES PEMBELAJARAN SUATU FUNGSI (MODEL) YANG MEMETAKAN SUATU ITEM DATA KEDALAM SATU KELAS DARI SEJUMLAH KELAS YANG TELAH DIDEFINISIKAN

DIGUNAKAN ALGORITMA KNN, SVM, DAN NAÏVE BAYES.

kNN ALGORITHM

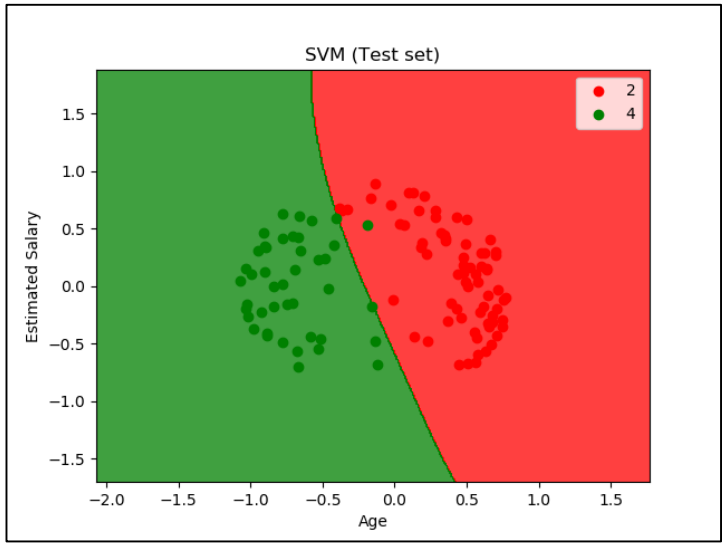


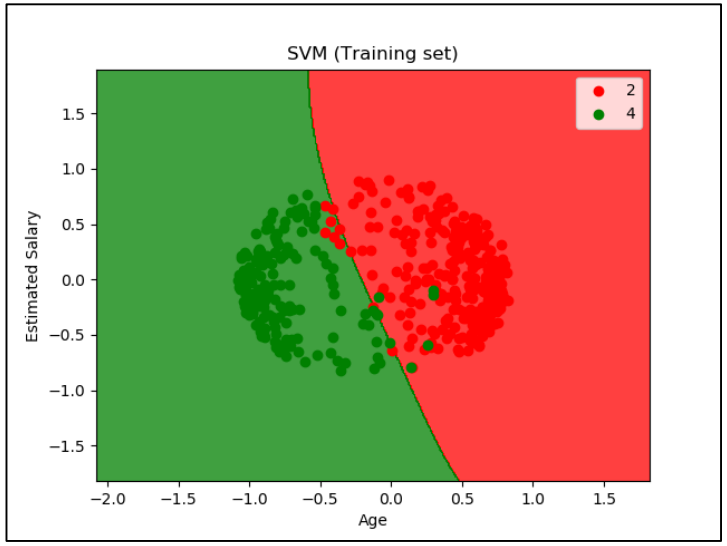


Test Accuracy: 0.9736842105263158 ≈ 97.37%

Train Accuracy: 0.9736263736263736 ≈ 97.36%

SUPPORT VECTOR MACHINE

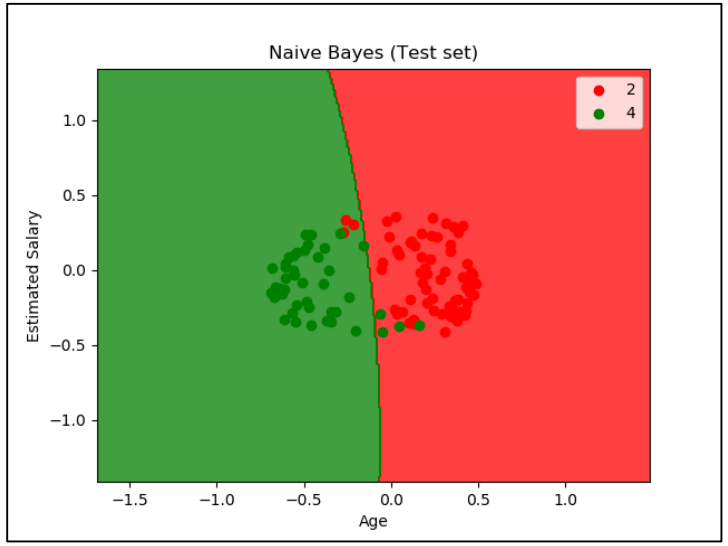


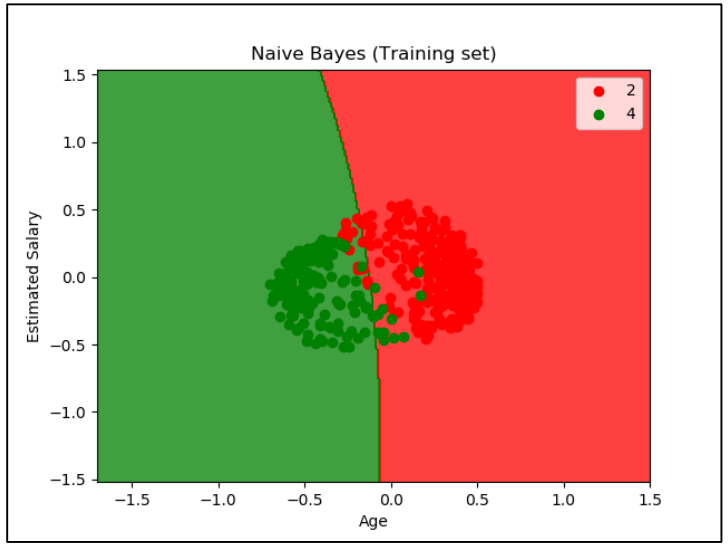


Test Accuracy: 0.9912280701754386 ≈ 99.12%

Train Accuracy: 0.9692307692307692 ≈ 96.92%

NAÏVE BAYES ALGORITHM





Test Accuracy: 0.9385964912280702 ≈ 93.85%

Train Accuracy: 0.9318681318681319 ≈ 93.19%

PERFORMANCE ANALYSIS (precision, recall, f1 score, support)

kNN

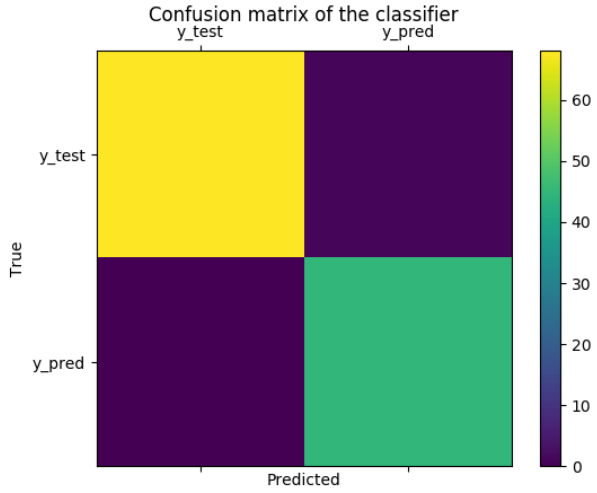
Type	Precision	recall	f1-score	support
2	1.00	0.96	0.98	70
4	0.94	1.00	0.97	44
micro avg	0.97	0.97	0.97	114
macro avg	0.97	0.98	0.97	114
weighted avg	0.98	0.97	0.97	114

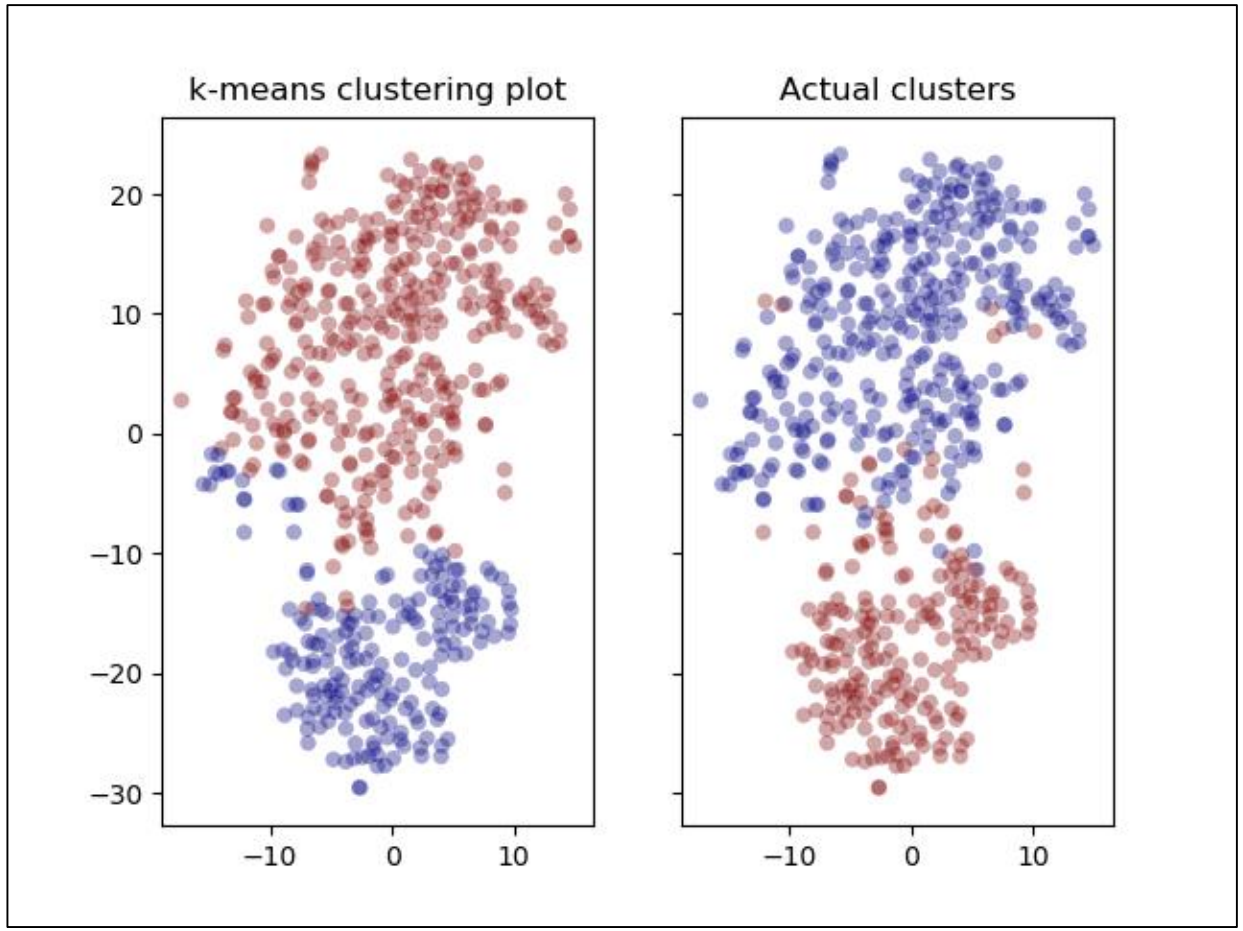
NaiveBayes

Type	Precision	recall	f1-score	support
2	0.98	0.91	0.95	70
4	0.88	0.98	0.92	44
micro avg	0.94	0.94	0.94	114
macro avg	0.93	0.95	0.94	114
weighted avg	0.94	0.94	0.94	114

SVM

Type	Precision	recall	f1-score	support
2	1.00	0.93	0.96	70
4	0.90	1.00	0.95	44
micro avg	0.96	0.96	0.97	114
macro avg	0.97	0.96	0.95	114
weighted avg	0.96	0.96	0.96	114

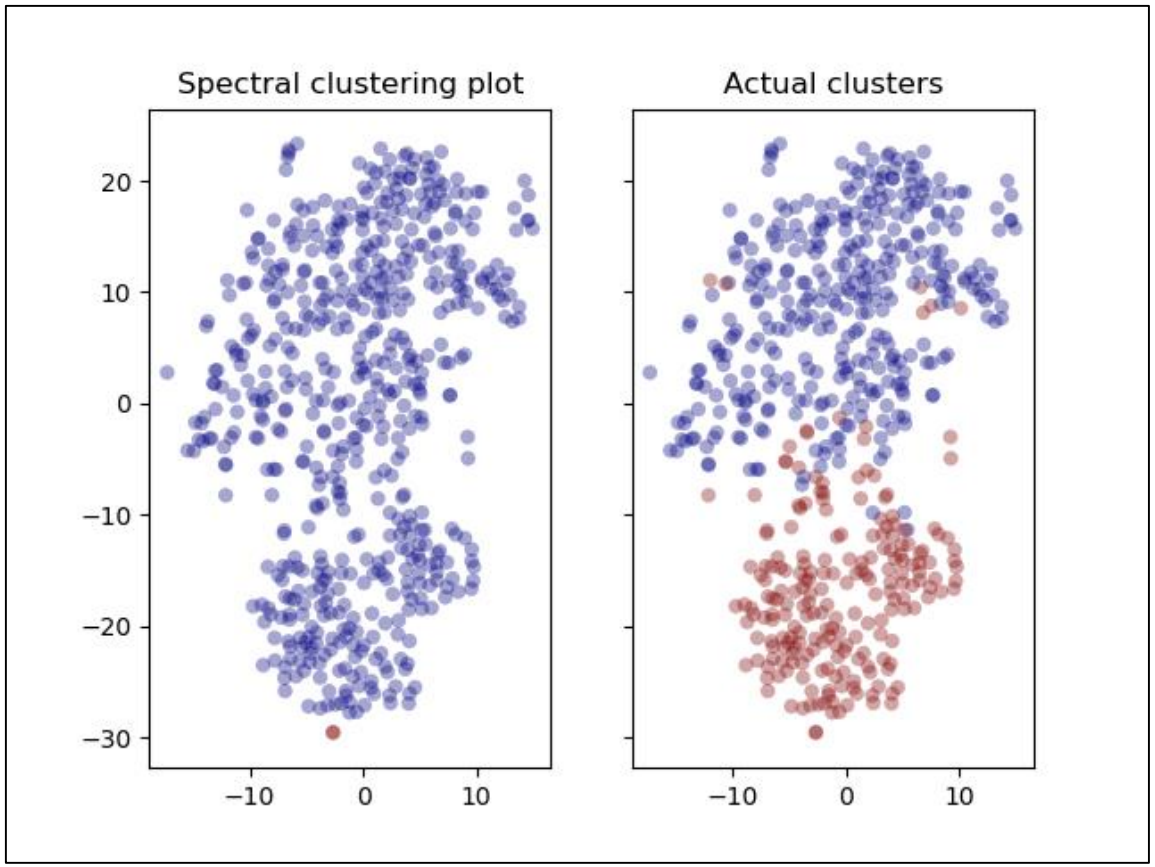


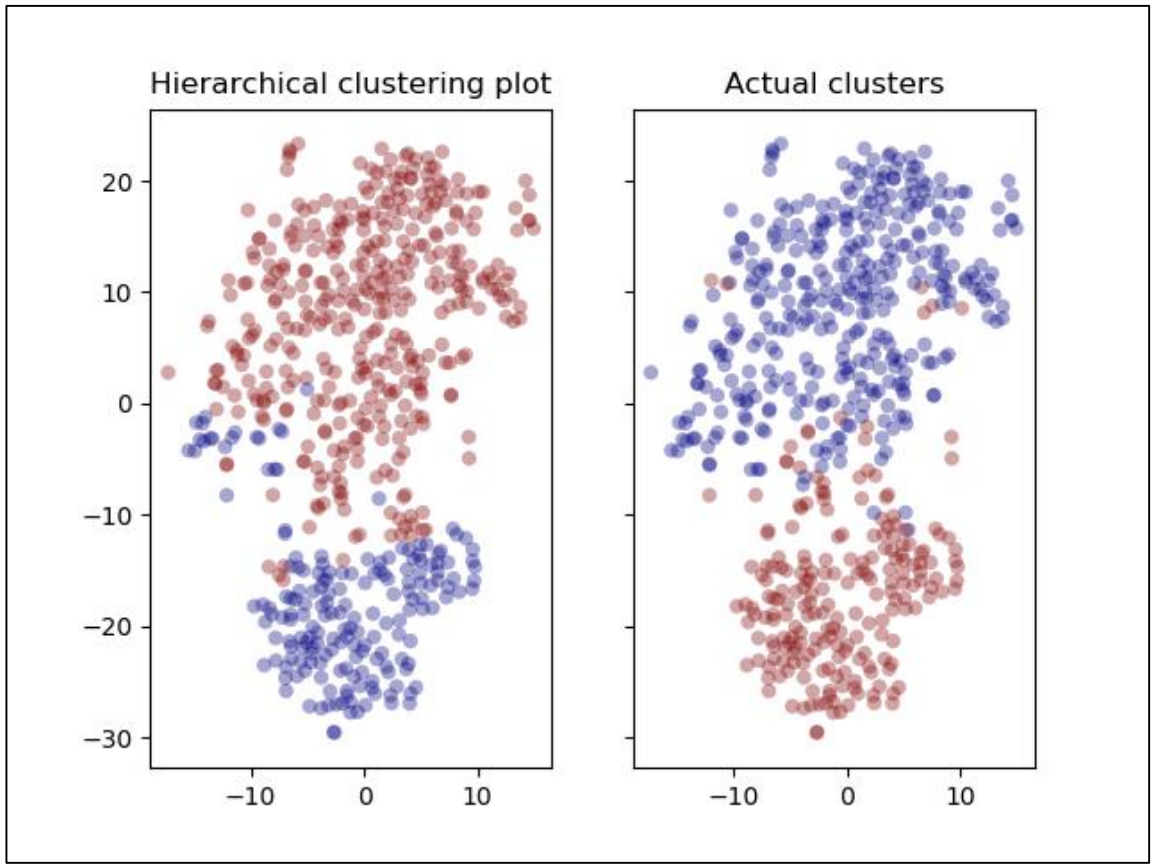


CLUSTERING

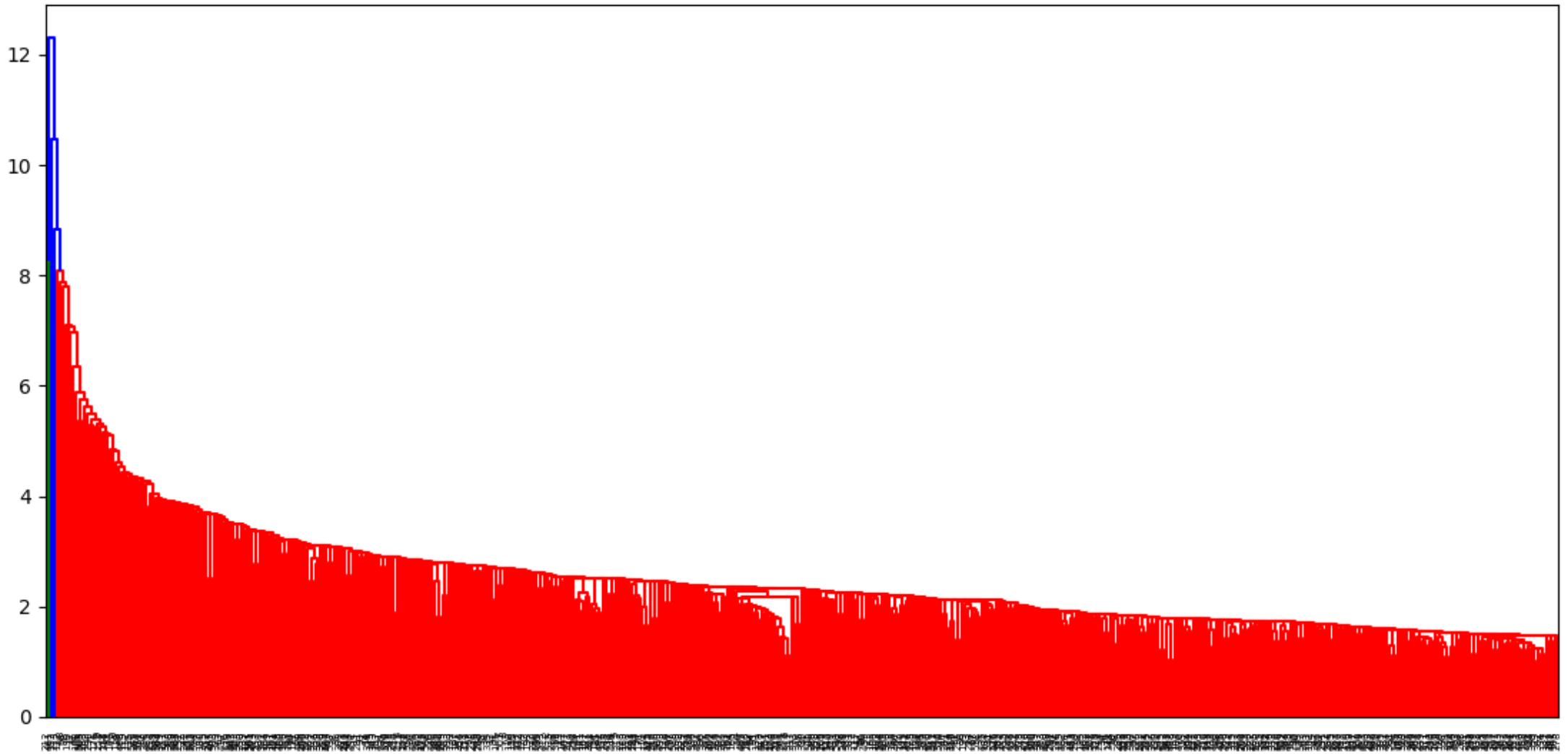
IDENTIFIKASI DAN PENGELOMPOKAN DARI CLASS (YANG DISEBUT JUGA DENGAN CLUSTER/GROUP) UNTUK SUATU HIMPUNAN OBYEK SEDEMIKIAN HINGGA ANGGOTA DARI SUATU CLUSTER SEDAPAT MUNGKIN MEMPUNYAI SIFAT YANG MIRIP DENGAN SESAMA ANGGOTA CLUSTER.

ALGORITMA YANG DIGUNAKAN ADALAH K-MEANS, SPECTRAL CLUSTERING, HIERARCHICAL CLUSTERING


Spectral clustering plot
Actual clusters
20
10
0
−10
−20
−30
−10
0
10
−10
0
10


Hierarchical clustering plot
Actual clusters
20
10
0
−10
−20
−30
−10
0
10
−10
0
10

DENDOGRAM

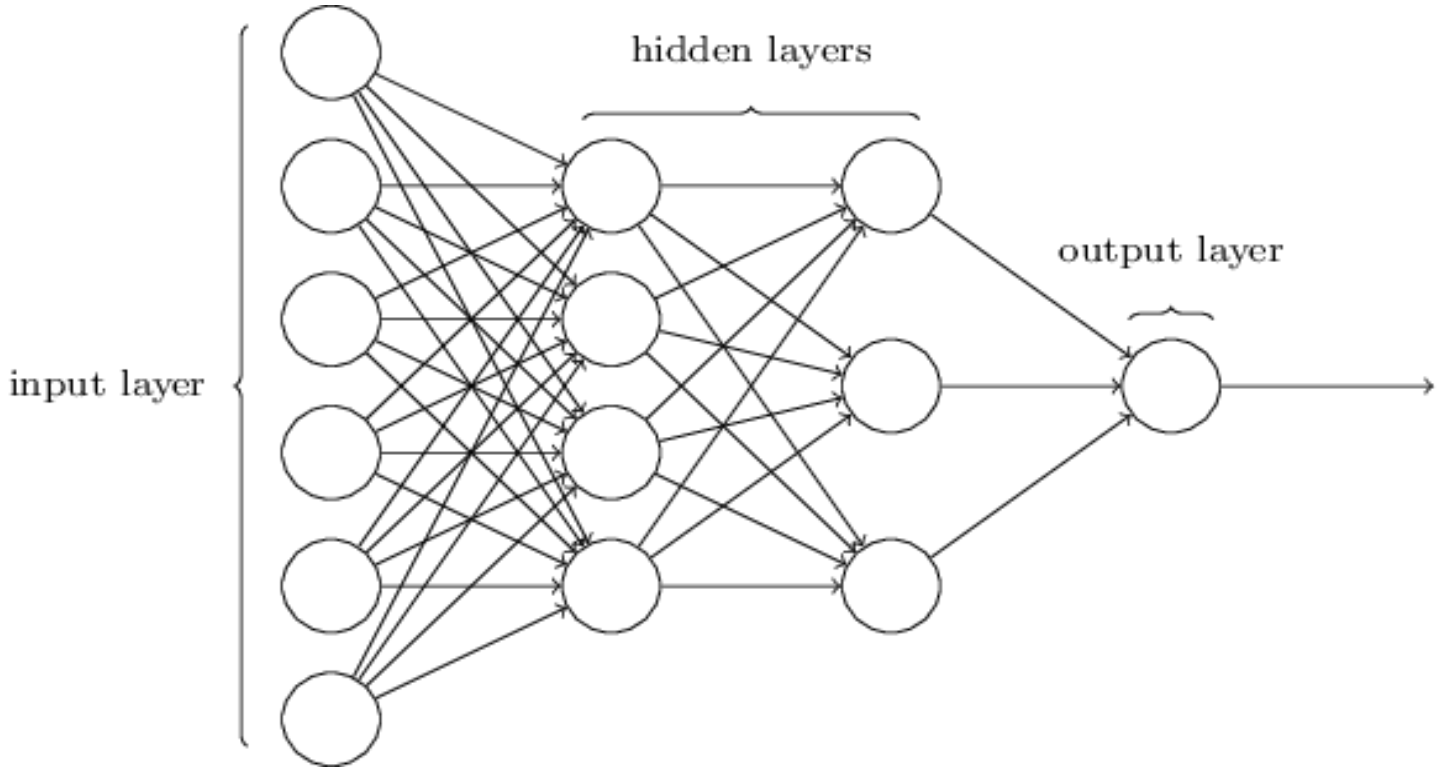


SEQUENTIAL PATTERN

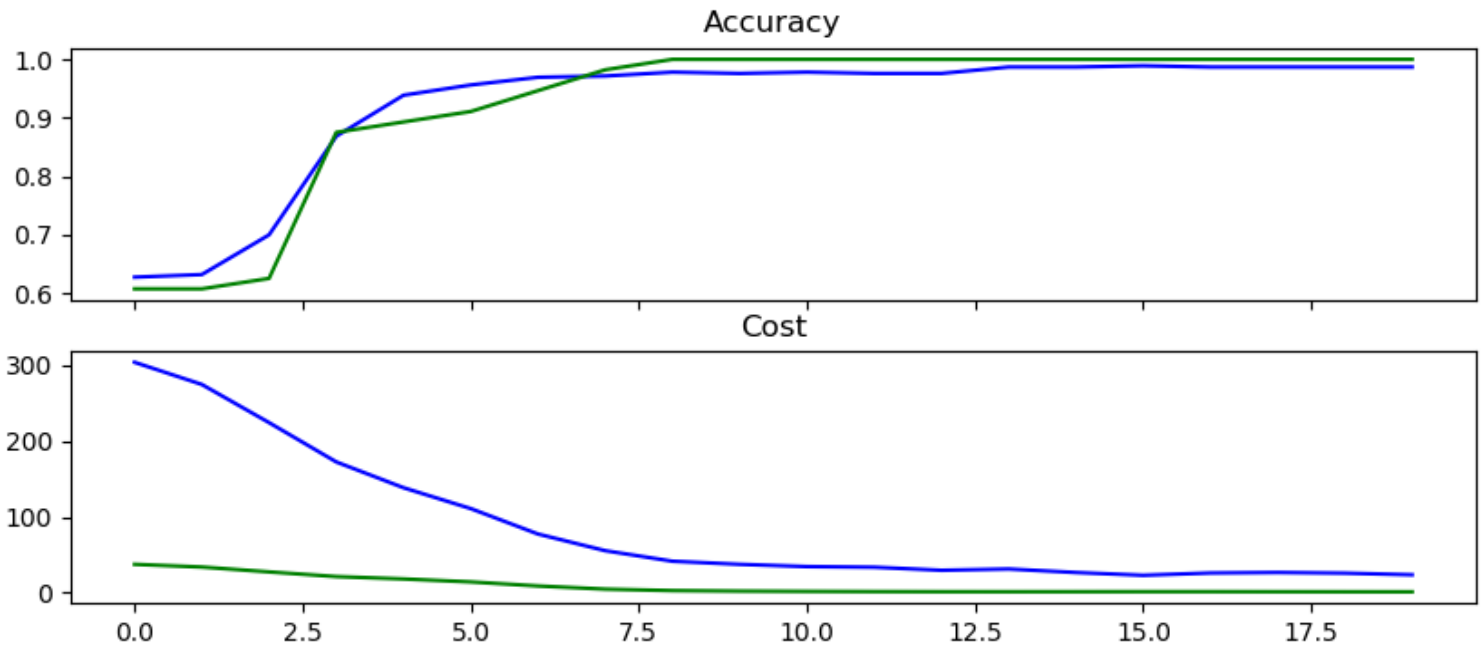
TEKNIK DALAM DATA MINING UNTUK MEMPEROLEH POLA PADA SUATU BARISAN DATA.

PREDIKSI DENGAN MULTI LAYER PERCEPTRON

PADA DASARNYA, MLP ADALAH *PERCEPTRON* YANG MEMILIKI *LAYER* ATAU LAPISAN TAMBAHAN DIANTARA *LAYER* INPUT (NEURON X_i) DAN LAYER OUTPUT (NEURON Y_i) YANG DISEBUT DENGAN *HIDDEN LAYER*. PROSES PERHITUNGAN DARI SETIAP NEURONNYA SAMA DENGAN *PERCEPTRON*. SINYAL OUTPUT NEURON (v) DIMASUKKAN KEDALAM SEBUAH FUNGSI AKTIVASI. (FAUSETT, 2006)(HAM & KOSTANIC, 2001).



HASIL PREDIKSI




```
# Neural Network Parameters
learning_rate = 0.005
training_dropout = 0.9
display_step = 1
batch_size = 100
accuracy_history = []
cost_history = []
valid_accuracy_history = []
valid_cost_history = []
```


THANK YOU
VENANSIUSRT | SVMIHAR