Data Testing dan training

TOOL AND TECHNIQUES FOR COMPUTATIONAL ANALYSIS

Dr. Muhammad Said Hasibuan

Materi diberikan Tgl 8 mei 2021

Target Pembelajaran

- Mahasiswa memiliki pemahaman tentang Data Testing dan Data Training
- Mahasiswa dapat membuat rancangan Evaluasi
- Mahasiswa dapat membuat laporan tulisan ilmiah atau buku Data Mining

Agenda

- Data Training
- Data Testing
- Model Evaluasi
- Confusion Matrix

Metode Data Mining

- Estimation (Estimasi)
  - Linear Regression (LR), Neural Network (NN), Deep Learning (DL), Support Vector Machine (SVM), Generalized Linear Model (GLM) etc
- Forecasting (Prediksi/Peramalan)
  - Linear Regression (LR), Neural Network (NN), Support Vector Machine (SVM), Generalized Linear Model (GLM) etc
- Classification (Klasifikasi)
  - Decision Tree (CART, ID3,C4.5, Credal,Naïve Bayes, KNN, Linear Discrimination Analysis (LDA), Logistic Regresion
- Clustering (Klastering)
  - K-Means, K-Medodis, Self Organizing Map (SOM), Fuzzy C Means
- Association (Asosiasi)
  - FP-Growth, A Priori, Coefficient of Correlation, Chi Square

Tipe Data Set

- Data Nominal : Laki laki, perumpuan
- Data Numerik : 12, 13,54

Contoh Penerapan Algoritma Estimasi

Contoh Data

1. Estimasi Waktu Pengiriman Pizza

Customer	Jumlah Pesanan (P)	Jumlah Traffic Light (TL)	Jarak (J)	Waktu Tempuh (T)
1	3	3	3	16
2	1	7	4	20
3	2	4	6	18
4	4	6	8	36
...				
1000	2	4	2	12

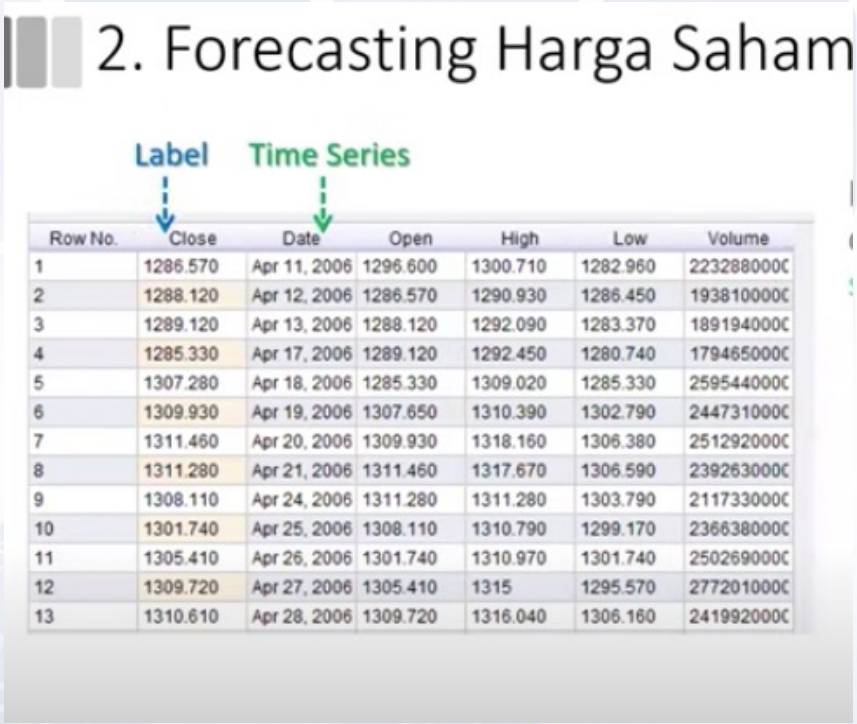
Penjelasan

- Data Numerik
- Atribut Numerik : jumlah pesanan, jumlah trafik, jarak
- Class Numerik : waktu tempuh

- Metode : Regresi Linear
- Hasil estimasi dapat berupa formula.

Contoh Penerapan Algoritma Forecasting

Contoh Data



2. Forecasting Harga Saham

Label Time Series

Row No.	Close	Date	Open	High	Low	Volume
1	1286.570	Apr 11, 2006	1296.600	1300.710	1282.960	223288000(
2	1288.120	Apr 12, 2006	1286.570	1290.930	1286.450	193810000(
3	1289.120	Apr 13, 2006	1288.120	1292.090	1283.370	189194000(
4	1285.330	Apr 17, 2006	1289.120	1292.450	1280.740	179465000(
5	1307.280	Apr 18, 2006	1285.330	1309.020	1285.330	259544000(
6	1309.930	Apr 19, 2006	1307.650	1310.390	1302.790	244731000(
7	1311.460	Apr 20, 2006	1309.930	1318.160	1306.380	251292000(
8	1311.280	Apr 21, 2006	1311.460	1317.670	1306.590	239263000(
9	1308.110	Apr 24, 2006	1311.280	1311.280	1303.790	211733000(
10	1301.740	Apr 25, 2006	1308.110	1310.790	1299.170	236638000(
11	1305.410	Apr 26, 2006	1301.740	1310.970	1301.740	250269000(
12	1309.720	Apr 27, 2006	1305.410	1315	1295.570	277201000(
13	1310.610	Apr 28, 2006	1309.720	1316.040	1306.160	241992000(

Penjelasan

- Data atribut/variabel numerik
- Class numerik
- Ada time series

Contoh Penerapan Algoritma Clasifikasi

Contoh Data

Label

NIM	Gender	Nilai UN	Asal Sekolah	IPS1	IPS2	IPS3	IPS 4	...	Lulus Tepat Waktu
10001	L	28	SMAN 2	3.3	3.6	2.89	2.9		Ya
10002	P	27	SMA DK	4.0	3.2	3.8	3.7		Tidak
10003	P	24	SMAN 1	2.7	3.4	4.0	3.5		Tidak
10004	L	26.4	SMAN 3	3.2	2.7	3.6	3.4		Ya
...									
...									
11000	L	23.4	SMAN 5	3.3	2.8	3.1	3.2		Ya

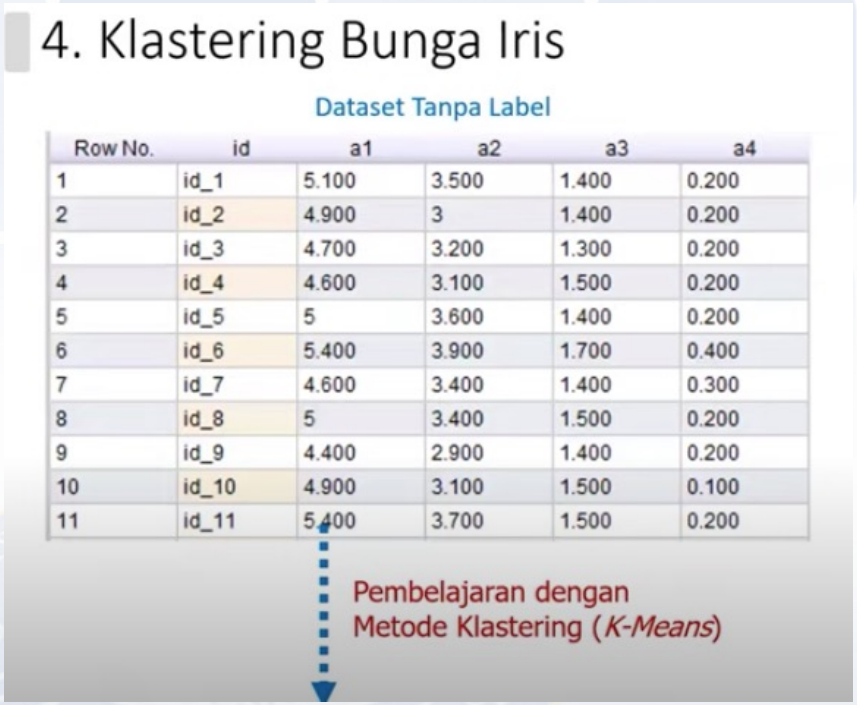
Pembelajaran dengan Metode Klasifikasi (*C4.5*)

Penjelasan

- Atribut boleh nominal atau numerik
- Label / class : nominal

Contoh Penerapan Algoritma Clustering

Contoh Data



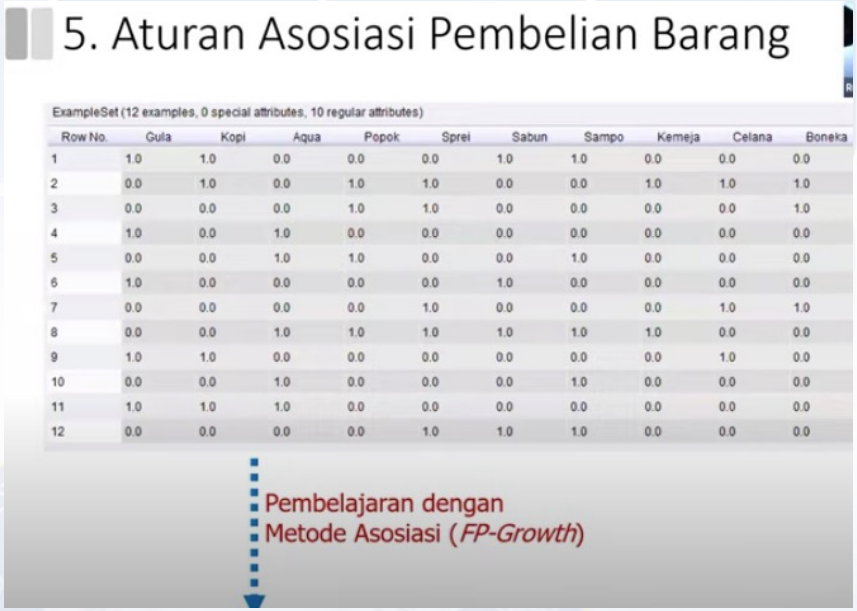
Row No.	id	a1	a2	a3	a4
1	id_1	5.100	3.500	1.400	0.200
2	id_2	4.900	3	1.400	0.200
3	id_3	4.700	3.200	1.300	0.200
4	id_4	4.600	3.100	1.500	0.200
5	id_5	5	3.600	1.400	0.200
6	id_6	5.400	3.900	1.700	0.400
7	id_7	4.600	3.400	1.400	0.300
8	id_8	5	3.400	1.500	0.200
9	id_9	4.400	2.900	1.400	0.200
10	id_10	4.900	3.100	1.500	0.100
11	id_11	5.400	3.700	1.500	0.200

Penjelasan

- Tidak ada label

Contoh Penerapan Algoritma Asosiasi

Contoh Data



Row No.	Gula	Kopi	Aqua	Popok	Sprei	Sabun	Sampo	Kemeja	Celana	Boneka
1	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0
2	0.0	1.0	0.0	1.0	1.0	0.0	0.0	1.0	1.0	1.0
3	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0
6	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0
8	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0	0.0
9	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
10	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
11	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0

Penjelasan

- Tidak ada hubungan dengan label atau class
- Apakah atribut masing masing saling berhubungan contoh Gula apakah ada hubungan dengan kopi dstnya

Data Training

- **Data training** digunakan untuk melatih algoritma.
- Contoh :

  Data Raport Siswa

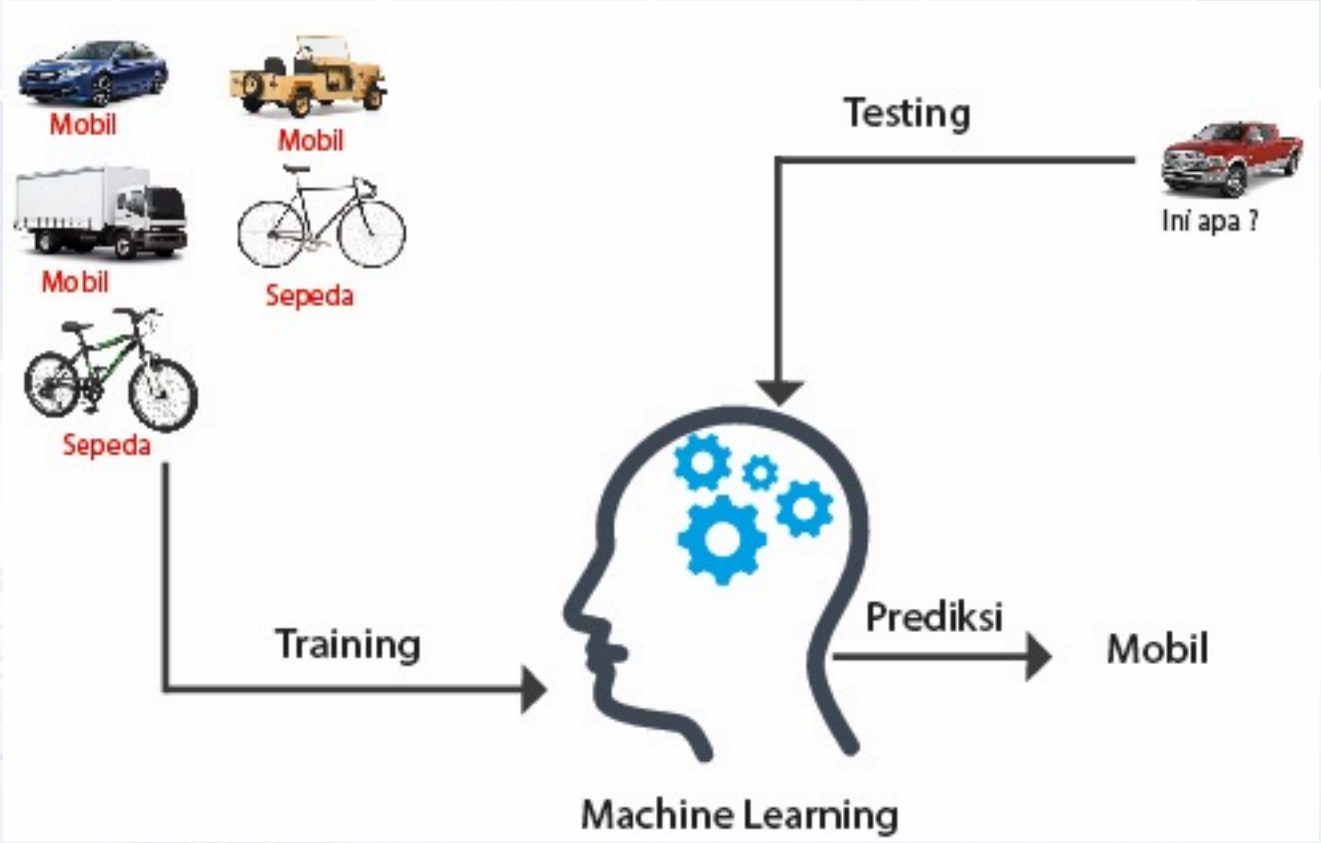
  Data wajah Manusia

  Data buah buahan

Data Testing

- **data testing** dipakai untuk mengetahui performa algoritma yang sudah dilatih sebelumnya ketika menemukan **data** baru yang belum pernah dilihat sebelumnya.

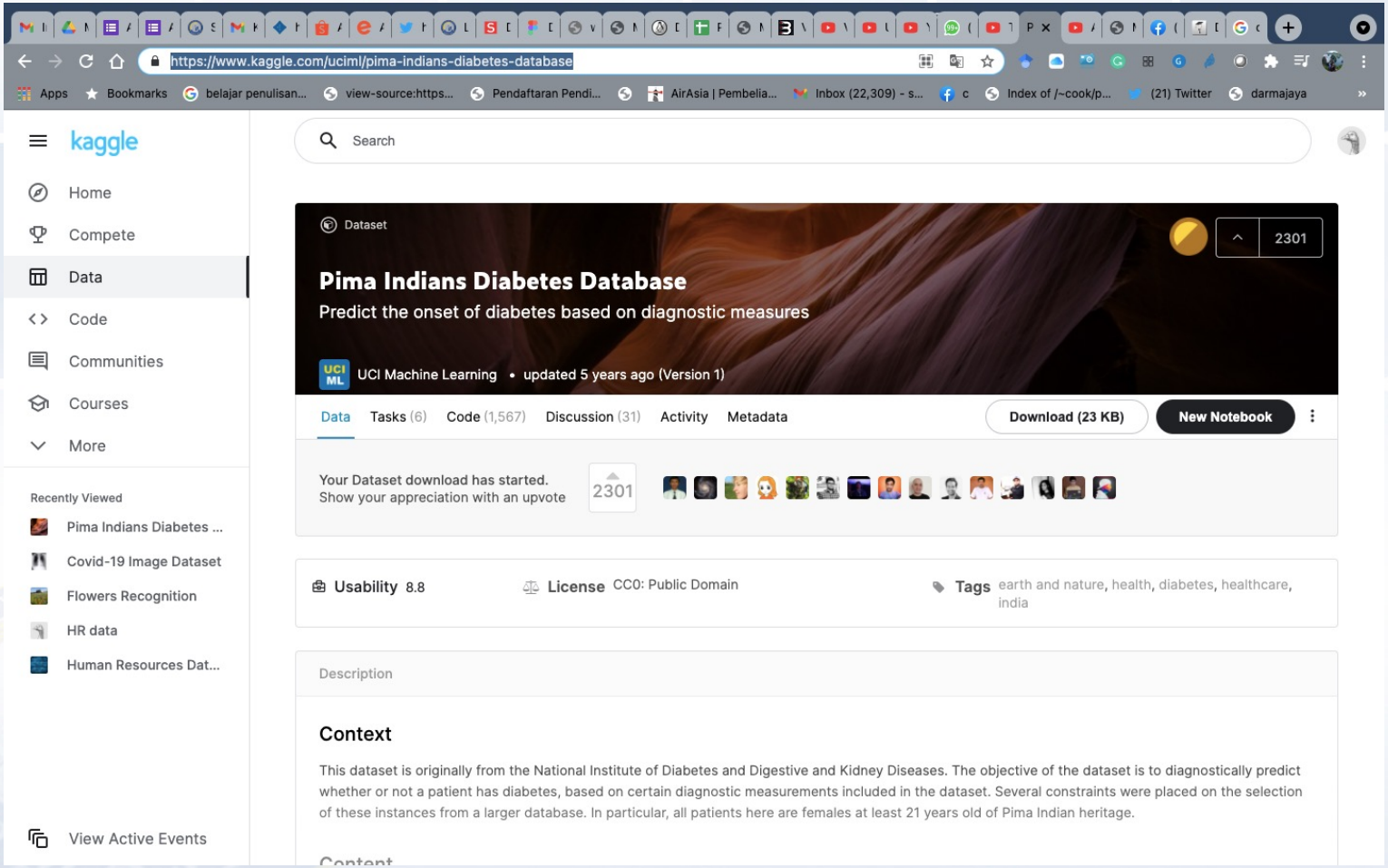
Konsep Awal


Mobil
Mobil
Mobil
Sepeda
Sepeda
Testing
Ini apa ?
Training
Prediksi
Mobil
Machine Learning

Validasi


A
Training
Test
Single Dataset
B
Training
Validation
Test
Single Dataset

Dataset


https://www.kaggle.com/uciml/pima-indians-diabetes-database
kaggle
Search
Home
Compete
Data
Code
Communities
Courses
More
Recently Viewed
Pima Indians Diabetes ...
Covid-19 Image Dataset
Flowers Recognition
HR data
Human Resources Dat...
View Active Events
Dataset
2301
Pima Indians Diabetes Database
Predict the onset of diabetes based on diagnostic measures
UCI ML
UCI Machine Learning • updated 5 years ago (Version 1)
Data
Tasks (6)
Code (1,567)
Discussion (31)
Activity
Metadata
Download (23 KB)
New Notebook
Your Dataset download has started.
Show your appreciation with an upvote
2301
Usability 8.8
License CC0: Public Domain
Tags earth and nature, health, diabetes, healthcare, india
Description
Context
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

Dataset

- Dataset dibagi 2 :

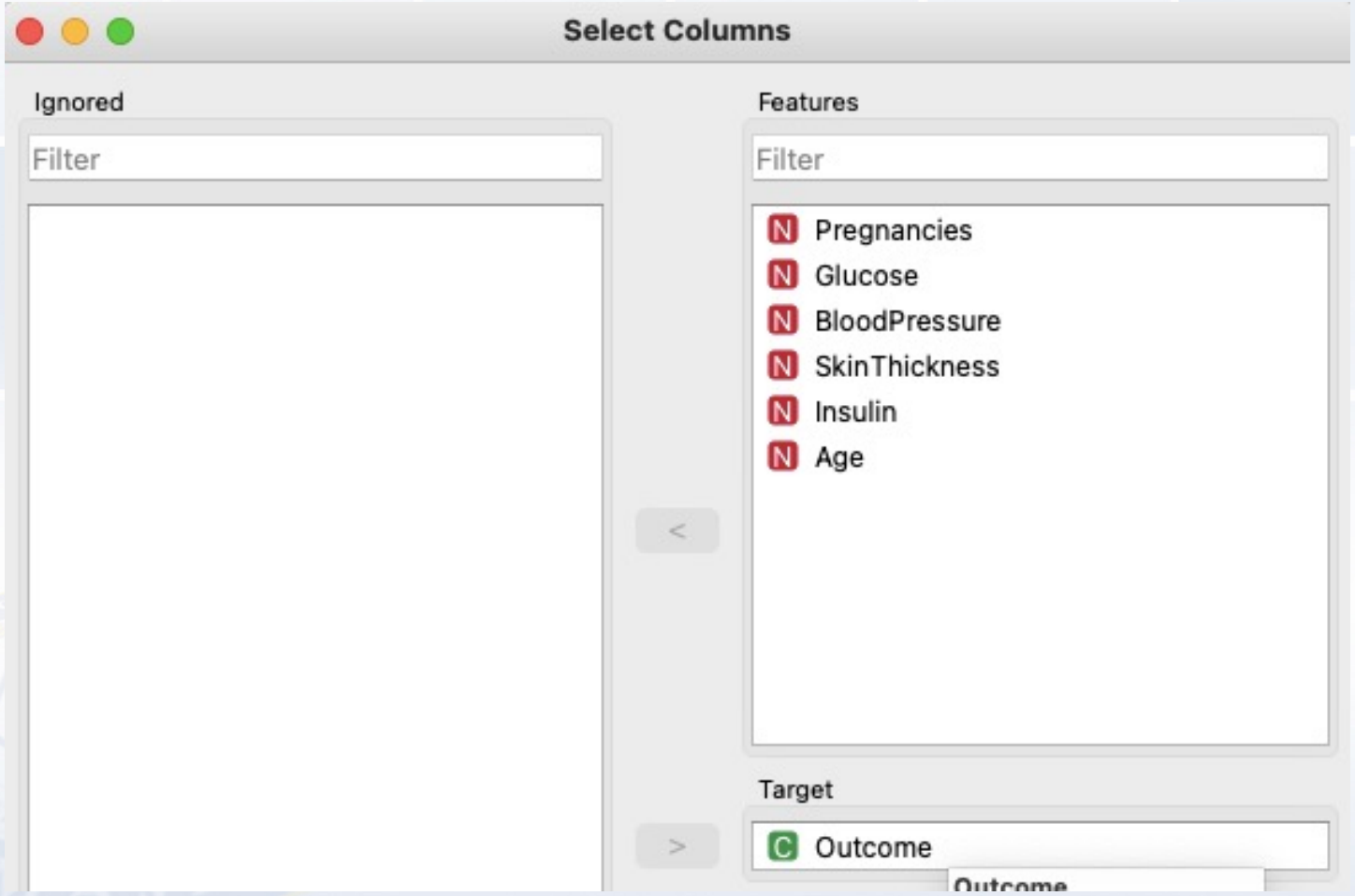
  Data Training

Pregnancies	Glucose	BloodPressui	SkinThicknes	Insulin	BMI	DiabetesPed	Age	Outcome
6	148	72	35	0	33.06.00	0,43541667	50	1
1	85	66	29	0	26.06.00	0,24375	31	0
8	183	64	0	0	23.03	0,46666667	32	1
1	89	66	23	94	28.01.00	0,11597222	21	0
0	137	40	35	168	43.01.00	2.288	33	1
5	116	74	0	0	25.06.00	0,13958333	30	0
3	78	50	32	88	31	0,17222222	26	1
10	115	0	0	0	35.03.00	0,09305556	29	0
2	197	70	45	543	30.05.00	0,10972222	53	1
8	125	96	0	0	0	0,16111111	54	1
4	110	92	0	0	37.06.00	0,13263889	30	0
10	168	74	0	0	38	0,37291667	34	1
10	139	80	0	0	27.01.00	1.441	57	0
1	189	60	23	846	30.01.00	0,27638889	59	1
5	166	72	19	175	25.08.00	0,40763889	51	1
7	100	0	0	0	30	0,33611111	32	1
0	118	84	47	230	45.08.00	0,38263889	31	1
7	107	74	0	0	29.06.00	0,17638889	31	1

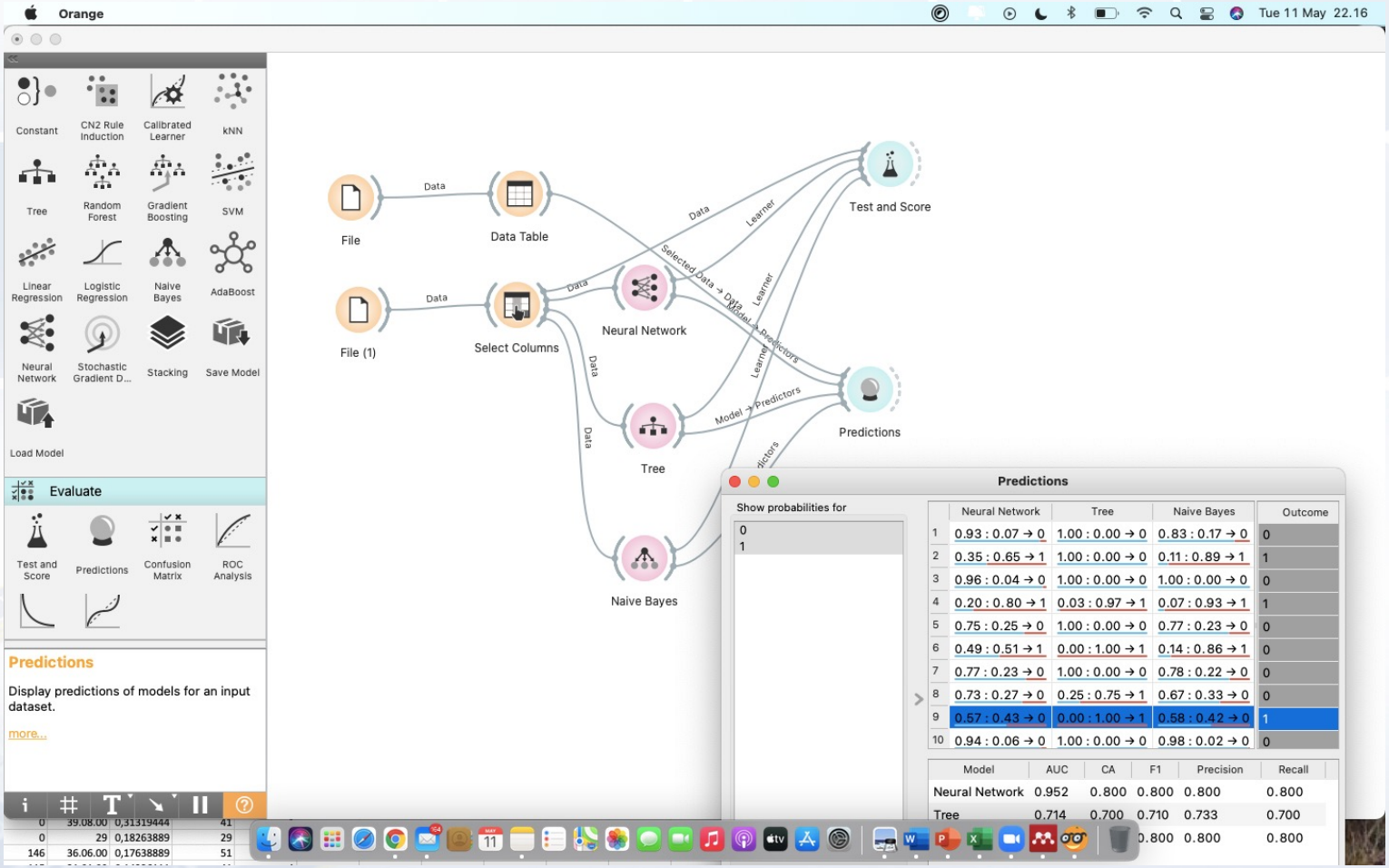
Data testing

Pregnancies	Glucose	BloodPressui	SkinThicknes	Insulin	BMI	DiabetesPed	Age
1	106	76	0	0	37.05.00	0,13680556	26
6	190	92	0	0	35.05.00	0,19305556	66
2	88	58	26	16	28.04.00	0,53194444	22
9	170	74	31	0	44	0,27986111	43
9	89	62	0	0	22.05	0,09861111	33
10	101	76	48	180	32.09.00	0,11875	63
2	122	70	27	0	36.08.00	00.34	27
5	121	72	23	112	26.02.00	0,17013889	30
1	126	60	0	0	30.01.00	0,24236111	47
1	93	70	31	0	30.04.00	0,21875	23

Setting di Data training dengan select column untuk menentukan target yaitu outcome



Model Evaluasi



Terima Kasih