

Laporan Tugas Besar

Pembelajaran Mesin

Tahap 2 : Classification

Kaenova Mahendra Auditama **Moh Adi Ikfini Mahfud**
1301190324 **1301194160**

CII3C3-IF-43-02



Pendahuluan

Tugas Besar pada Mata Kuliah Pembelajaran Mesin (CII3C3-IF-43-02) Classification merupakan tugas kedua dari dua proyek tugas yang ada. Pada tugas ini, kami diminta untuk melakukan klasifikasi terhadap dataset yang disediakan.

Permasalahan

Pada kasus ini kami diberikan suatu dataset terkait ketertarikan pelanggan dengan beberapa atribut-atribut. Data yang diberikan sebesar **285.831 records**. Adapun beberapa atribut seperti *id*, *Jenis_Kelamin*, *Umur*, *SIM*, *Kode_Daerah*, *Sudah_Asuransi*, *Umur_Kendaraan*, *Kendaraan_Rusak*, *Premi*, *Kanal_Penjualan*, *Lama_Berlangganan*, dan *Tertarik*.

Dengan data tersebut, kami diminta untuk melakukan klasifikasi terhadap orang yang tertarik dan yang tidak tertarik. Kami membenchmark beberapa model diantaranya Naive Bayes, Decision Tree, dan Artificial Neural Network

Eksplorasi dan Pra-Pemrosesan Data

```
[ ] 1 df_raw = pd.read_csv("https://raw.githubusercontent.com/kaenova/Malin_Tubes1/main/data/raw/kendaraan_train.csv")
    2 df_raw.head()
```

	id	Jenis_Kelamin	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik
0	1	Wanita	30.0	1.0	33.0	1.0	< 1 Tahun	Tidak	28029.0	152.0	97.0	0
1	2	Pria	48.0	1.0	39.0	0.0	> 2 Tahun	Pernah	25800.0	29.0	158.0	0
2	3	NaN	21.0	1.0	46.0	1.0	< 1 Tahun	Tidak	32733.0	160.0	119.0	0
3	4	Wanita	58.0	1.0	48.0	0.0	1-2 Tahun	Tidak	2630.0	124.0	63.0	0
4	5	Pria	50.0	1.0	35.0	0.0	> 2 Tahun	NaN	34857.0	88.0	194.0	0

```
[ ] 1 len(df_raw)
```

285831

Examples:

Total: 285831

Positive: 35006 (12.25% of total)

Label Balance pada Data Train

Examples:

Total: 47639

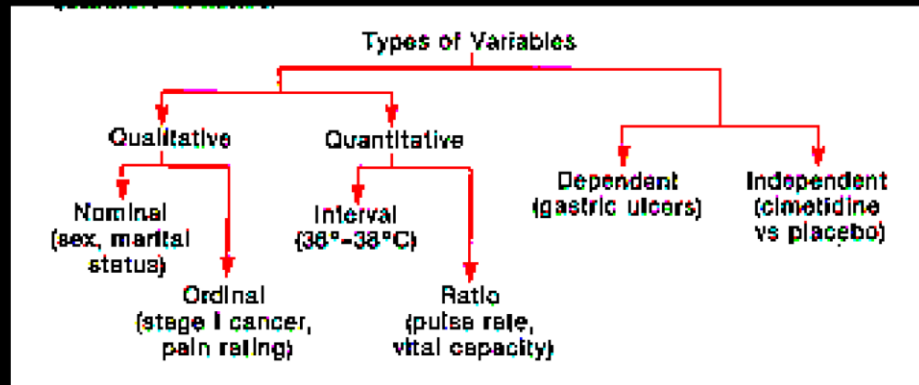
Positive: 5861 (12.30% of total)

Label Balance pada Data Test

Pertama, lakukan penghilangan data yang terlihat jelas sebagai data categorical seperti *Jenis_Kelamin*, *SIM*, *Sudah_Asuransi*, *Umur_Kendaraan*, dan *Kendaraan_Rusak*. Data yang diberikan tidak balance, tetapi memiliki rasio yang sama.

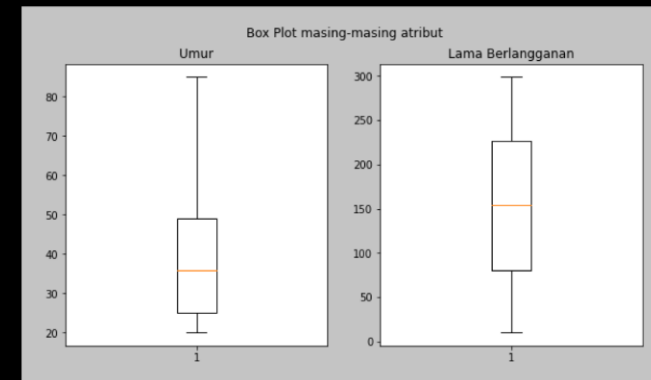
Kedua, melakukan penghilangan *records* jika pada salah satu data ada yang kosong atau *NaN*

Eksplorasi dan Pra-Pemrosesan Data (cont.)



Jenis Data

Eksplorasi Pada Jenis Data Quantitative yang Pasti



Boxplot Data Numerical

Terlihat Normal, tidak ada outlier.

Eksplorasi dan Pra-Pemrosesan Data (cont.)

Ekplorasi Pada Atribut yang Tidak Pasti

Premi

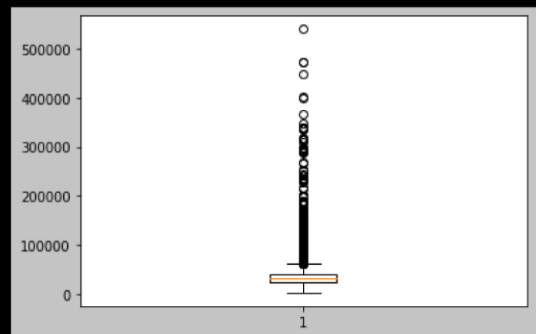
Menurut KBBI:

Dalam KBBI (<https://kbbi.kemdikbud.go.id/entri/premi>) kita bisa menganggap atribut ini merupakan quantitative

n hadiah (uang dan sebagainya) yang diberikan sebagai perangsang untuk meningkatkan prestasi kerja

n hadiah (dalam undian, perlombaan, pembelian)

n jumlah uang yang harus dibayarkan pada waktu tertentu kepada asuransi sosial: -- asuransi



Boxplot Atribut Premi

Total data: 171068
Not Outliers data: 85534

Jika dilakukan penghilangan data Outlier

Tidak digunakan sebagai fitur

Eksplorasi dan Pra-Pemrosesan Data (cont.)

Eksplorasi Pada Atribut yang Tidak Pasti

Kanal Penjualan

Bisa kita sinonimkan sebagai “Salur Penjualan”
Kalau dari kata-katanya tidak mungkin ini nilai quantitative,
yang sangat memungkinkan ini merupakan qualitative dalam bentuk numerical dan tidak bersifat ordinal.

```
[ ] 1 train_kanal = np.sort(df_train_analyze_kanal.astype('int').unique())
    2 test_kanal = np.sort(df_test_analyze_kanal.unique())
    3 all_kanal = set(train_kanal.tolist()+ test_kanal.tolist())

[ ] 1 not_in_train = []
    2 not_in_test = []
    3 for kanal in all_kanal:
    4     if kanal not in train_kanal:
    5         not_in_train.append(kanal)
    6     if kanal not in test_kanal:
    7         not_in_test.append(kanal)
    8
    9 print(not_in_train)
   10 print(not_in_test)
   11 len(not_in_test)
```

```
[ ]
[2, 6, 17, 27, 28, 50, 57, 69, 71, 74, 75, 82, 84, 100, 101, 110, 117, 123, 134, 137, 144]
21
```

Ada beberapa data yang tidak ada dalam data train ataupun data test
Ini harus diperhatikan ketika kita ingin melakukan One Hot Encoding.
Karena dalam One Hot Encoding bentuk data harus sama

Eksplorasi dan Pra-Pemrosesan Data (cont.)

Korelasi Terhadap Data yang Sudah Diketahui

	Umur	Lama_Berlangganan
Umur	1.000000	-0.001032
Lama_Berlangganan	-0.001032	1.000000

Terhadap Data Numerikal

	Jenis_Kelamin	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Kanal_Penjualan
Jenis_Kelamin	1.00	0.0	0.01	0.01	0.03	0.01	0.04
SIM	0.00	1.0	0.00	0.00	0.00	0.00	0.00
Kode_Daerah	0.01	0.0	1.00	0.06	0.09	0.05	0.02
Sudah_Asuransi	0.01	0.0	0.06	1.00	0.15	0.68	0.20
Umur_Kendaraan	0.03	0.0	0.09	0.15	1.00	0.16	0.39
Kendaraan_Rusak	0.01	0.0	0.05	0.68	0.16	1.00	0.21
Kanal_Penjualan	0.04	0.0	0.02	0.20	0.39	0.21	1.00

Terhadap Data Kategorikal

Kendaraan_Rusak dan Sudah_Asuransi akan kita jadikan “Hipotesis Best Feature”

Eksplorasi dan Pra-Pemrosesan Data (cont.)

Tahapan PreProcessing Data

1. Menghapus atribut yang tidak dijadikan fitur: [id, Premi, Tertarik]
2. Menghilangkan Data Duplikat atau Data NULL
3. Melakukan Min-Max Scaling pada Atribut Numerical: "Umur" dan "Lama_Berlangganan"
4. Oversampling agar data sedikit balance menggunakan SMOTE
5. Melakukan One Hot Encoding pada Atribut Kategorikal:
['Jenis_Kelamin', 'SIM', 'Sudah_Asuransi', 'Umur_Kendaraan', 'Kendaraan_Rusak', 'Kanal_Penjualan']
Perlu diperhatikan kita juga harus menambahkan kolom secara manual pada data yang tidak ada.

Tahapan Preparation Data

Membuat Data Train dengan K-Fold Stratified untuk Mendapatkan Model Terbaik
Ketika train.

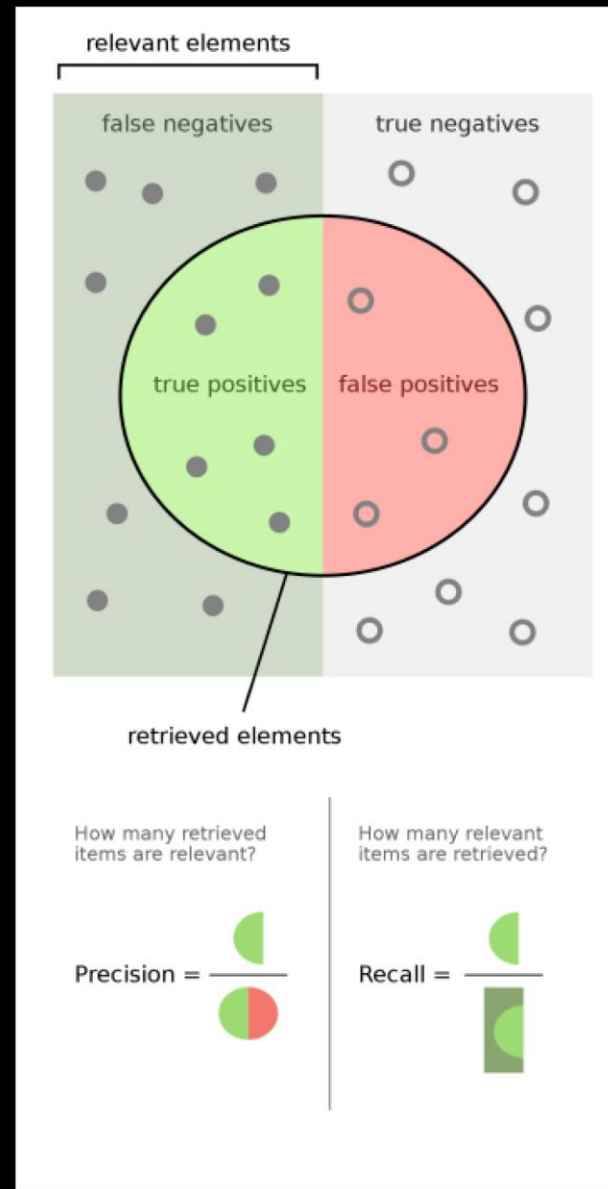
*Berdasarkan Ekplorasi Data

Pemodelan

Metrics yang digunakan

$$\text{Akurasi} = \frac{TP + FP}{TP + FP + TN + FN}$$

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{tp}{tp + \frac{1}{2}(fp + fn)}$$



Pemodelan

Model yang Digunakan

Naive Bayes

Decision Tree

**Artificial
Neural Netwok**

Pengimplementasian

Python dapat dilihat melalui link referensi:

https://kaenova-link.pages.dev/school/malin_tubes2

atau

<https://bit.ly/KaenovaMalinTubes2>


Hasil_(utama)

Model	Metrics					
	Akurasi	W. Precision	W. Recall	W. F1 Score	Precision (Tertarik)	Recall (Tertarik)
”Hyphothesis” Best Feature						
Decision Tree	0.64	0.99	0.64	0.70	0.25	0.98
Gaussian NB	0.64	0.99	0.64	0.70	0.25	0.98
ANN	0.64	0.99	0.64	0.70	0.25	0.98
All Feature						
Decision Tree	0.82	0.83	0.82	0.82	0.29	0.30
ANN	0.73	0.89	0.73	0.77	0.29	0.88



Table 1: Performa Data Testing pada setiap Model (NB = Naive Bayes, ANN = Artificial Neural Network, W. = Weighted)

Kesimpulan

Model **Naive Bayes** dapat melakukan klasifikasi pada dataset yang sudah disediakan. Dengan beberapa eksplorasi dan pra-pemrosesan kami menggunakan atribut “Kendaraan_Rusak” dan “Sudah_Asuransi” dan mendapatkan **F1 Score 0.70** dan **Akurasi 0.64**.

 kendaraan_test_predicted.csv

Open with ▾



	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Jenis_Kelamin	Umur	SIM	Kode_Daerah	Sudah_Asuransi	Umur_Kendaraan	Kendaraan_Rusak	Premi	Kanal_Penjualan	Lama_Berlangganan	Tertarik	Prediksi
2	0	Wanita	49	1	8	0	1-2 Tahun	Pemah	46963	26	145	0	
3	1	Pria	22	1	47	1	< 1 Tahun	Tidak	39624	152	241	0	
4	2	Pria	24	1	28	1	< 1 Tahun	Tidak	110479	152	62	0	
5	3	Pria	46	1	8	1	1-2 Tahun	Tidak	36266	124	34	0	
6	4	Pria	35	1	23	0	1-2 Tahun	Pemah	26963	152	229	0	
7	5	Pria	26	1	28	1	< 1 Tahun	Tidak	42721	152	198	0	
8	6	Wanita	24	1	28	1	< 1 Tahun	Tidak	65801	152	160	0	
9	7	Wanita	40	1	28	0	1-2 Tahun	Pemah	30981	26	79	0	
10	8	Pria	23	1	15	1	< 1 Tahun	Tidak	32365	152	219	0	
11	9	Wanita	43	1	28	0	1-2 Tahun	Pemah	65380	25	41	1	
12	10	Wanita	53	1	28	1	1-2 Tahun	Tidak	80184	26	30	0	
13	11	Pria	28	1	46	1	< 1 Tahun	Tidak	25657	152	133	0	
14	12	Pria	24	1	21	1	< 1 Tahun	Tidak	35817	152	35	0	
15	13	Wanita	21	1	22	1	< 1 Tahun	Tidak	29404	160	249	0	
16	14	Wanita	23	1	36	0	< 1 Tahun	Pemah	36266	152	235	0	
17	15	Pria	21	1	36	1	< 1 Tahun	Tidak	2630	152	93	0	
18	16	Wanita	22	1	21	1	< 1 Tahun	Tidak	44554	152	224	0	
19	17	Pria	22	1	4	1	< 1 Tahun	Tidak	41897	152	265	0	
20	18	Pria	50	1	46	0	1-2 Tahun	Pemah	32127	154	10	1	
21	19	Pria	56	1	28	0	> 2 Tahun	Pemah	53157	26	15	1	
22	20	Pria	75	1	12	1	1-2 Tahun	Tidak	22910	124	16	0	
23	21	Wanita	52	1	28	0	1-2 Tahun	Pemah	40796	122	187	1	
24	22	Pria	27	1	42	1	< 1 Tahun	Tidak	38495	26	127	0	

Referensi

(2021). Correlation coefficient: Simple definition, formula, easy steps.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cramer, H. (2021). *Mathematical Methods of Statistics*.

Hans, R. (2020). Jenis-jenis algoritma supervised machine learning.

Potdar, K., S., T., and D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, 175(4):7–9.