Komputasi dan Sains Data

Prediksi Pengambilan Asuransi pada Travel Insurance

(k-NN, Decision Tree)

Suri Dian Pratama (2006614771)
Gabriella Aileen Mendrofa (2106779472)
Nadhilah Farhana (2106779516)

Klasifikasi Pembelian Travel Insurance

Latar Belakang

Sebuah perusahaan tour & travel menawarkan paket asuransi perjalanan kepada pelanggannya.

Perusahaan tertarik untuk mengetahui faktor apa saja yang mempengaruhi pembelian paket asuransi.

Selain itu, perusahaan juga perlu mengetahui apakah pelanggan akan tertarik untuk membelinya berdasarkan sejarah basis datanya.

Tujuan Klasifikasi

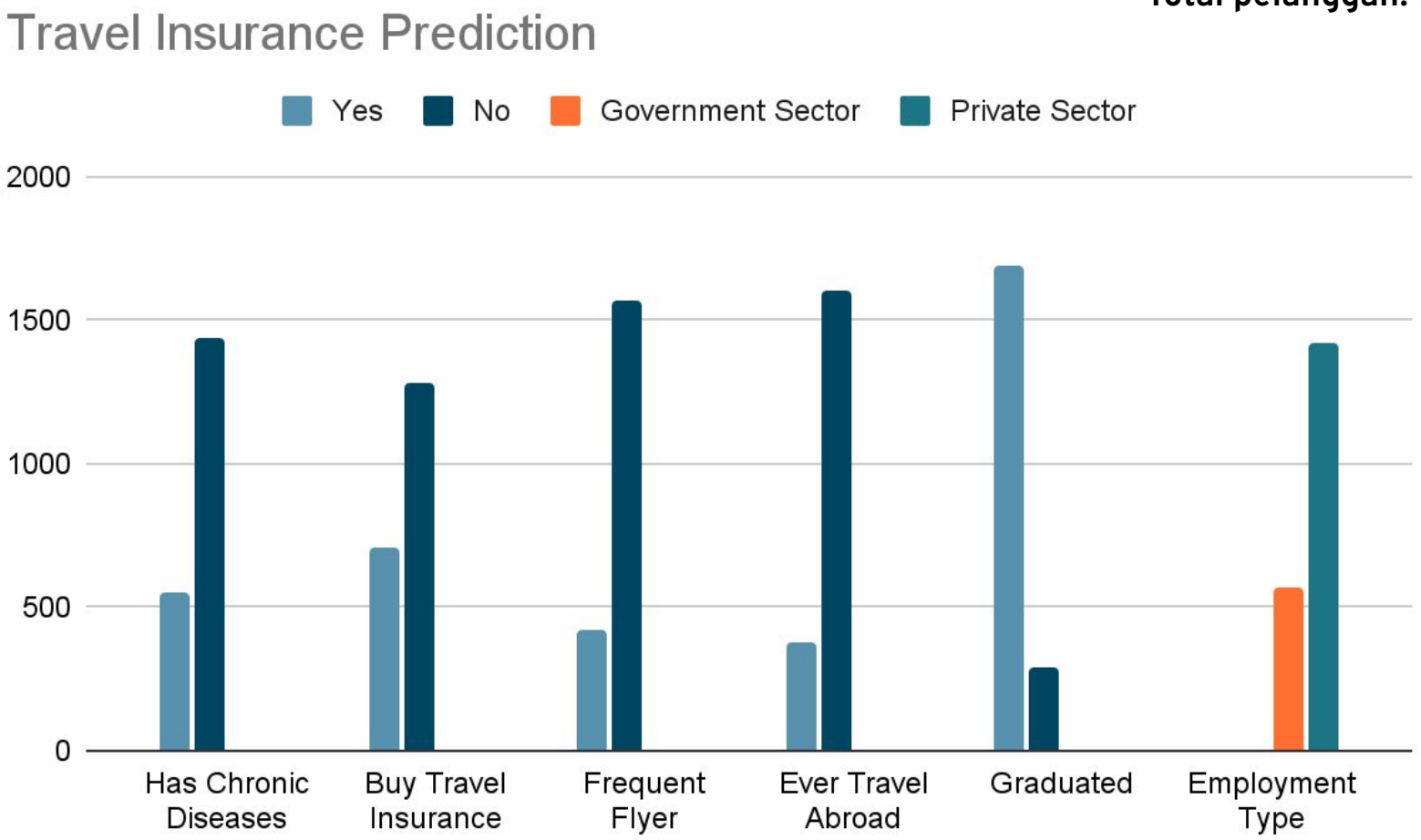
Menentukan **algoritma klasifikasi terbaik** di antara k-NN dan Decision Tree untuk mengelompokkan pelanggan ke dalam 2 kategori:
-Membeli travel insurance
-Tidak membeli travel insurance

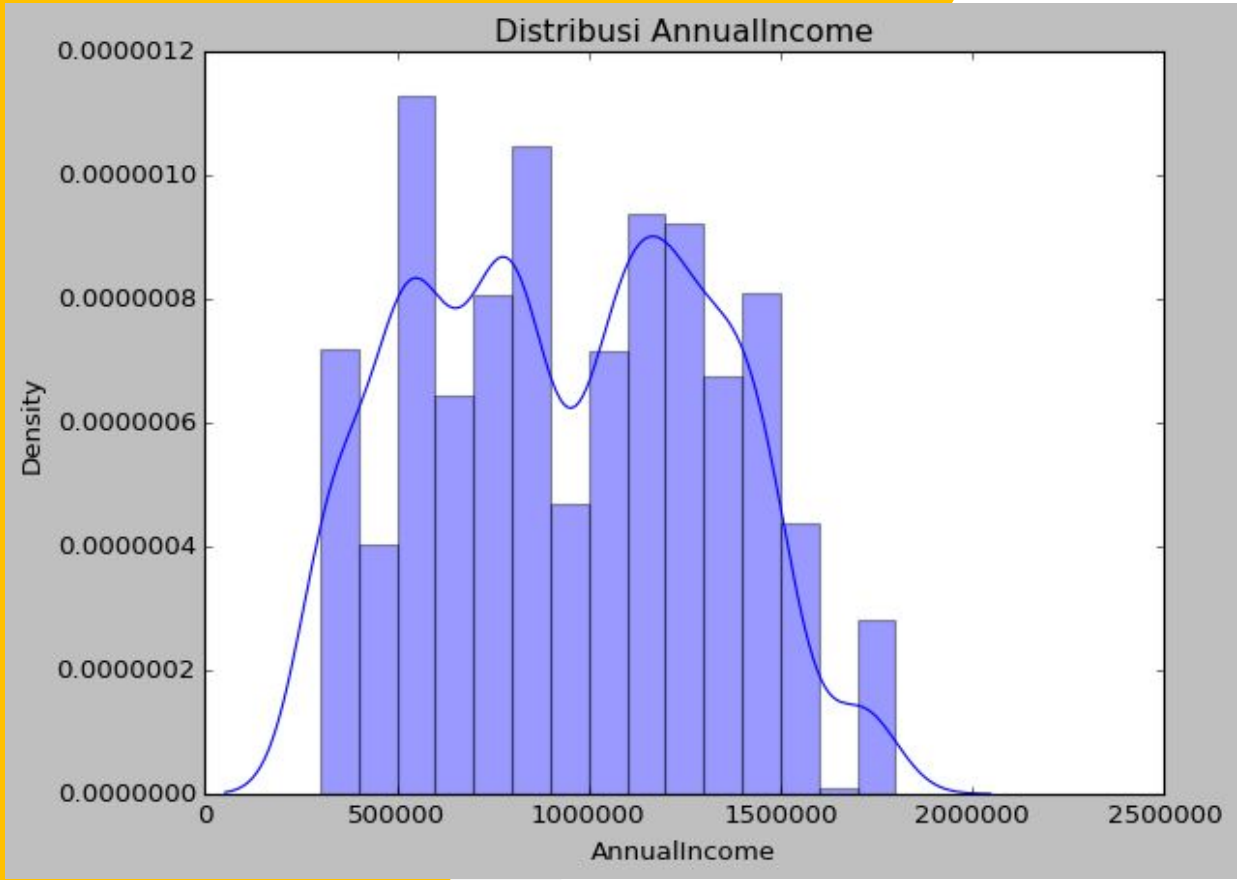
Variabel Data

- Age
- Employment Type
- GraduateOrNot
- AnnualIncome
- FamilyMembers
- ChronicDiseases
- FrequentFlyer
- EverTravelledAboard

Get the Data

Total pelanggan: 1987

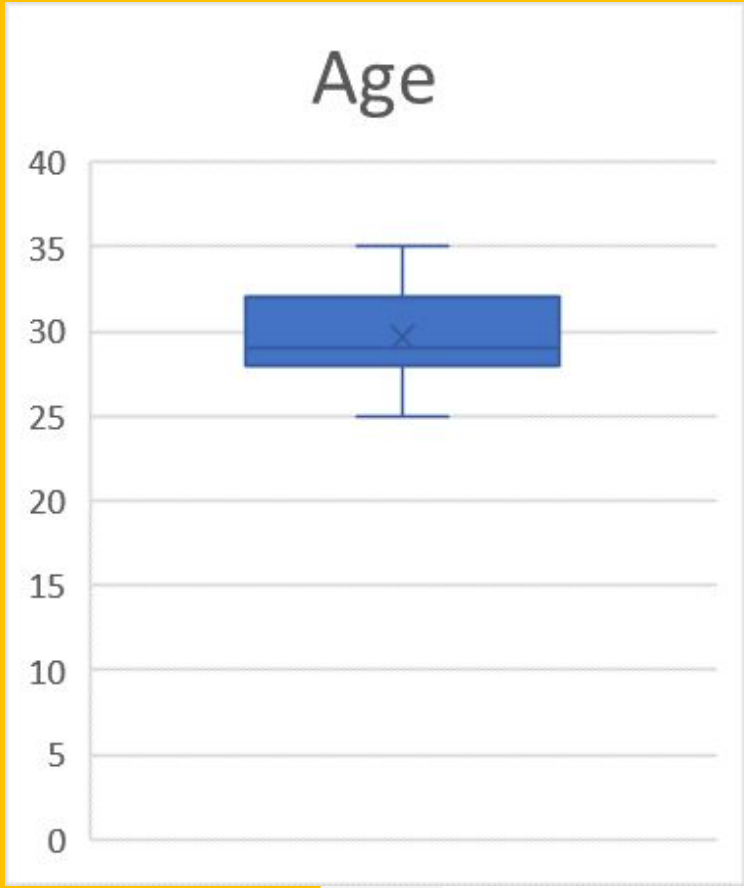




Annual Income

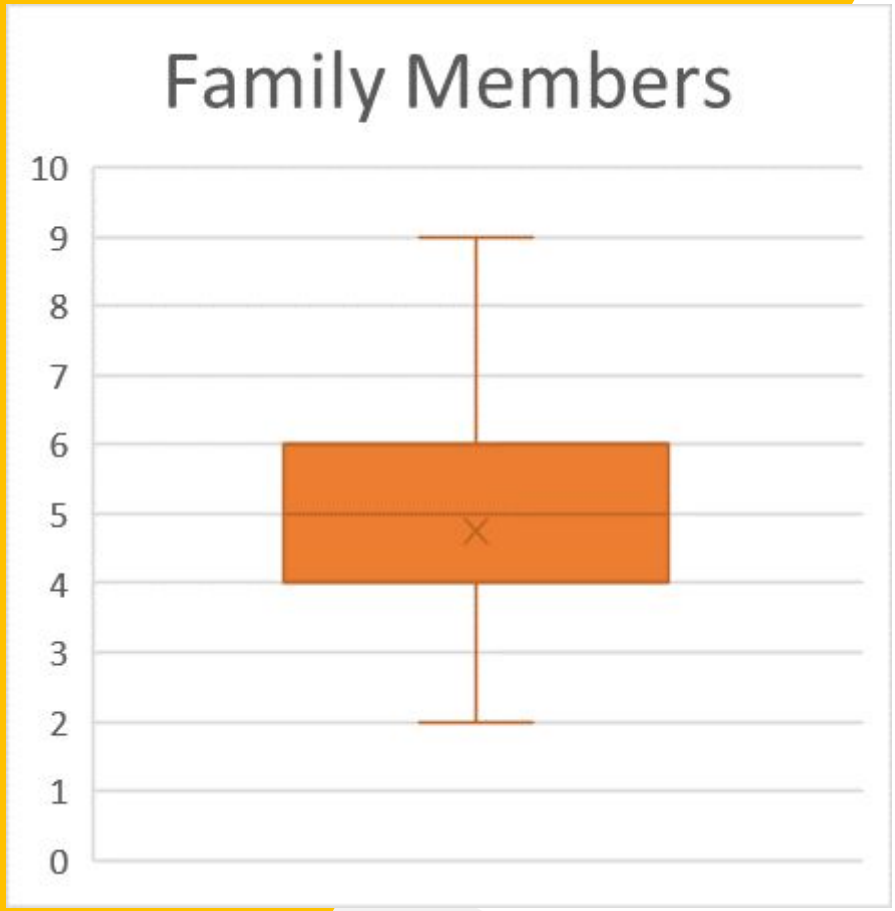
Annual Income tidak berdistribusi normal

Mayoritas pelanggan memiliki annual income pada interval 500.000 sampai 1.500.000



Age

Mayoritas pelanggan memiliki usia pada interval 28 sampai 32 tahun.



Family Members

Mayoritas pelanggan memiliki 4 sampai 6 anggota keluarga.

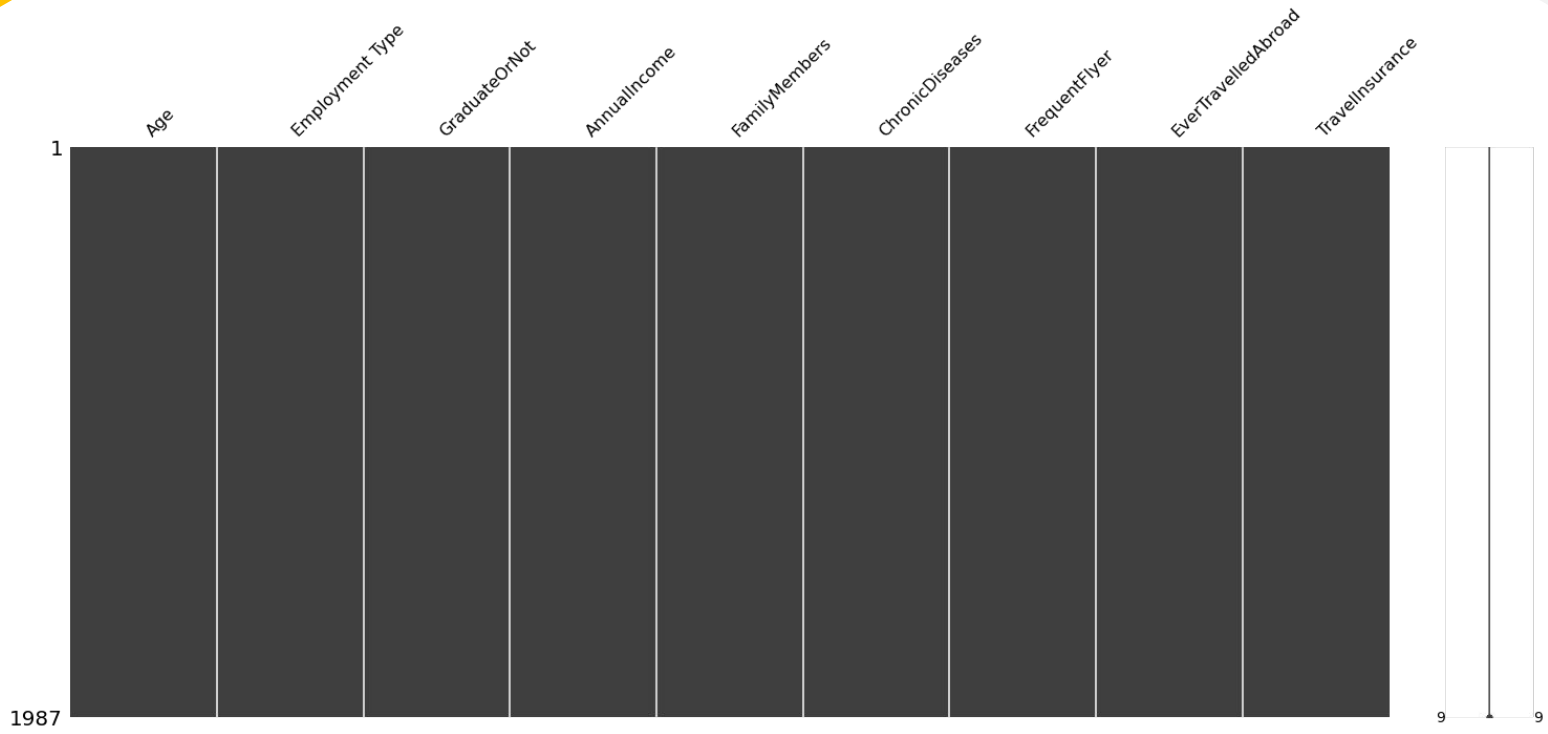
Preprocessing

Drop Kolom Unnamed: 0

Kolom Unnamed: 0 tidak digunakan dalam proses analisis dan klasifikasi, sehingga dihapus dari data

	Age	EmploymentType	GraduateOrNot	AnnualIncome	FamilyMembers	ChronicDiseases	FrequentFlyer	EverTravelledAbroad	TravelInsurance
1042	31	0	1	500000	2	0	0	0	0
1555	26	0	1	1400000	9	1	0	1	1
634	34	0	0	1300000	4	0	0	0	0
727	34	1	1	1100000	7	1	0	0	1
751	28	0	1	1350000	9	0	0	1	1

Heatmap Missing Value



Tidak terdapat missing value pada setiap variabel

Mengubah Tipe Data

Tipe Data Awal

Variabel	Tipe Data
Age	Integer
Annual Income	
Family Members	
Chronic Diseases	
Travel Insurance	
Frequent Flyer	Object
Ever Travelled Abroad	
Employment Type	
Graduate or Not	

Tipe Data Akhir

Variabel	Tipe Data
Age	Integer
Annual Income	
Family Members	
Chronic Diseases	Category
Travel Insurance	
Frequent Flyer	
Ever Travelled Abroad	
Employment Type	
Graduate or Not	

Statistik Deskriptif

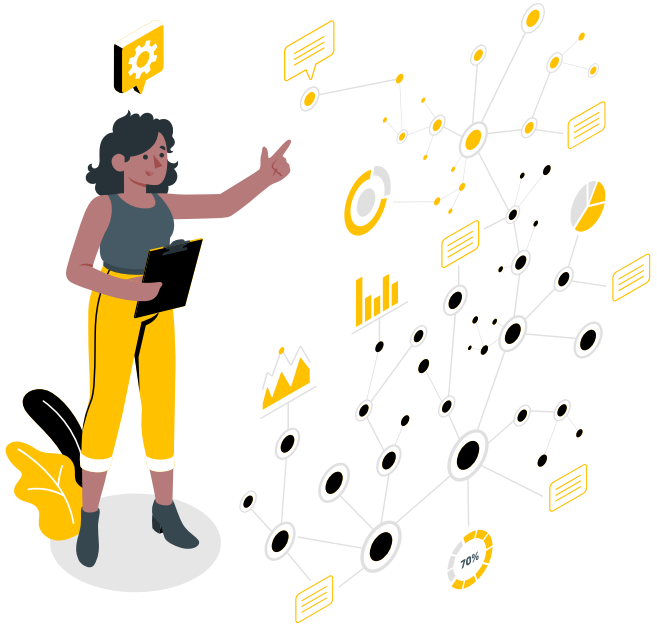
Numerik

Variabel	Age	Annual Income	Family Members
mean	29.65	932762.95	4.75
std	2.91	376855.68	1.6
min	25	300000	2
25%	28	600000	4
50%	29	900000	5
75%	32	1250000	6
max	35	1800000	9

Nilai Mean dan Median tidak berbeda jauh → kemungkinan data tidak skewed/tidak ada outlier

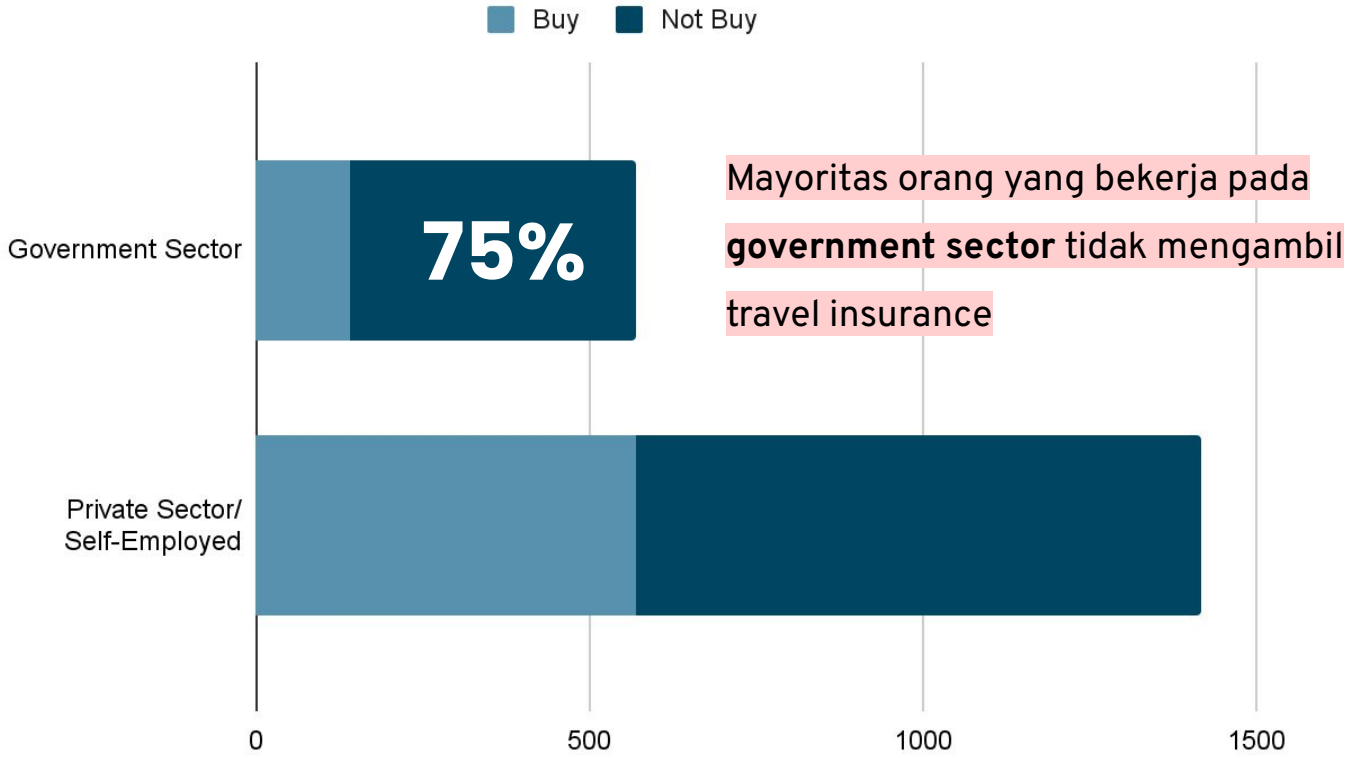
Nilai min dan max sesuai (tidak terdapat keanehan) → tidak terdeteksi anomali/noise

Visualisasi Data



Employment Type

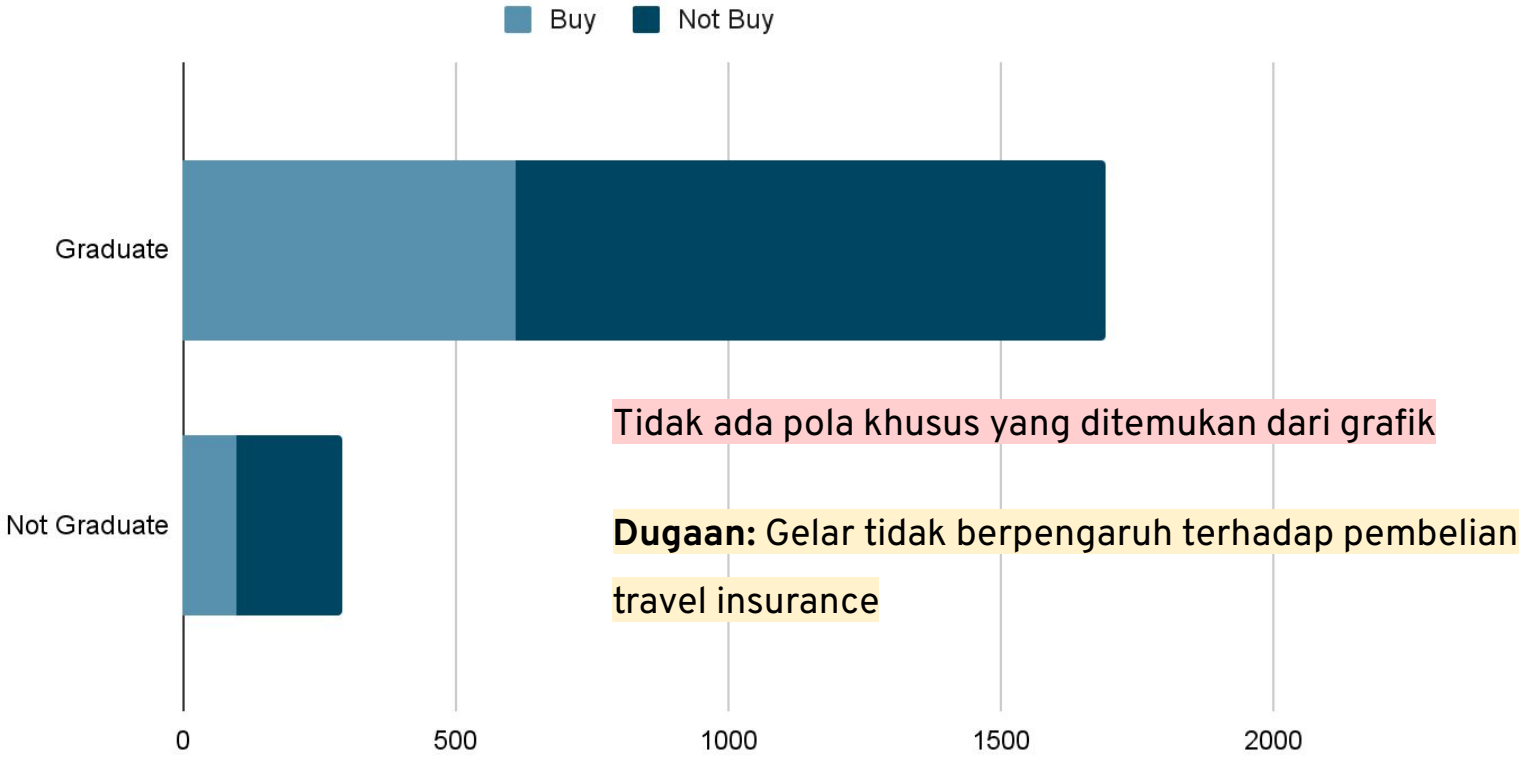
Total government sector: 570
Total private sector: 1417



Mayoritas orang yang bekerja pada **government sector** tidak mengambil travel insurance

Graduate Or Not

Total graduate: 1692
Total not graduate: 295

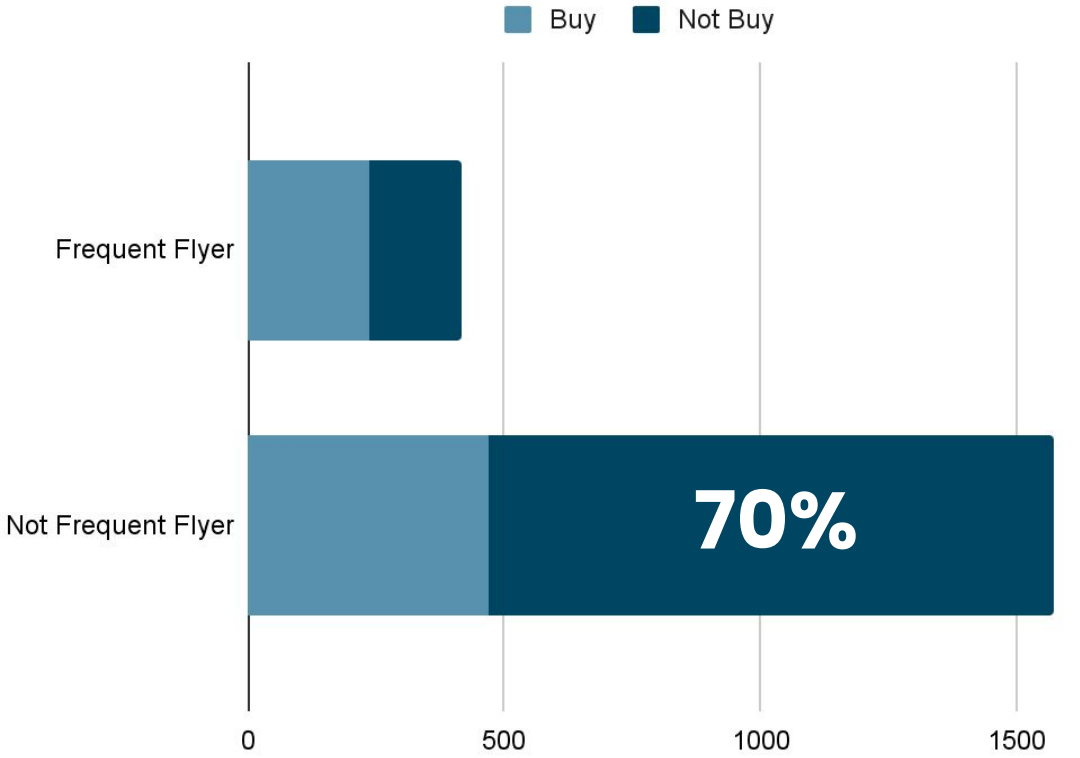


Tidak ada pola khusus yang ditemukan dari grafik

Dugaan: Gelar tidak berpengaruh terhadap pembelian travel insurance

Frequent Flyer

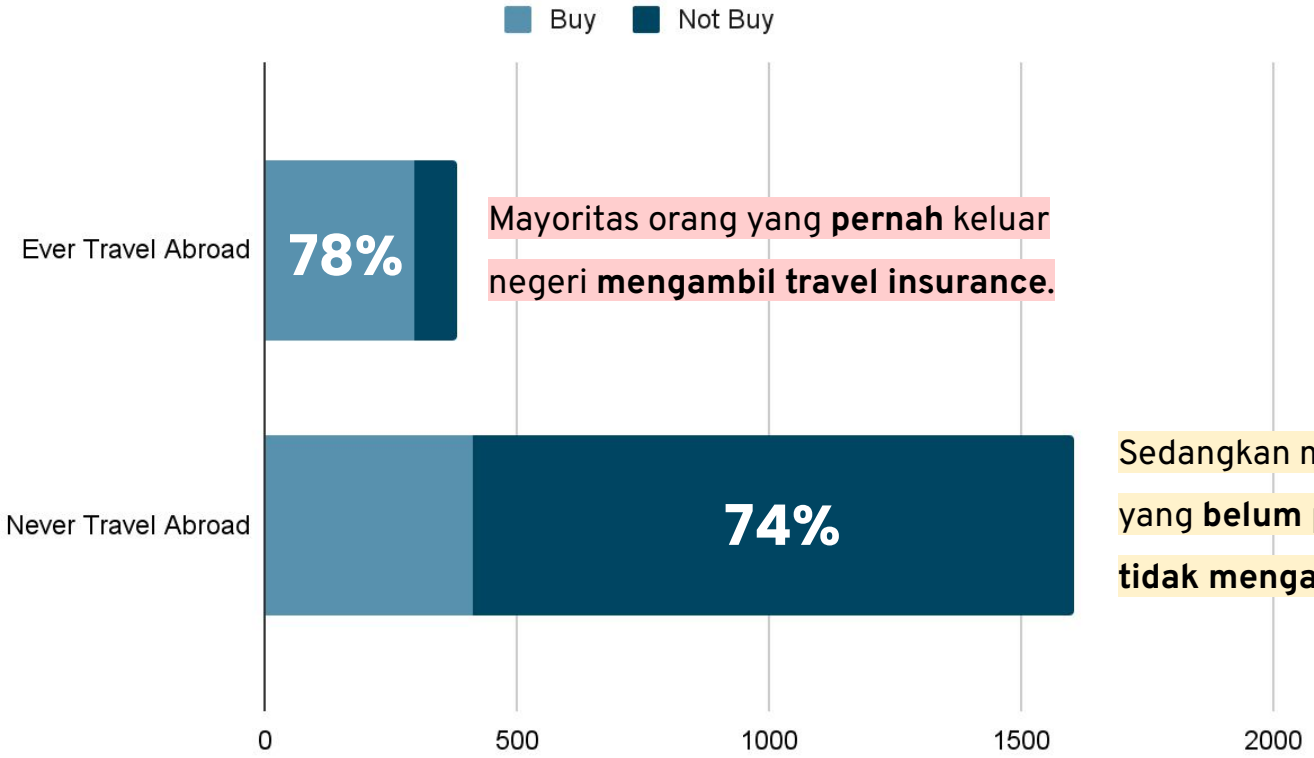
Total frequent flyer: 417
Total not frequent flyer: 1570



Mayoritas orang yang **tidak sering melakukan penerbangan** tidak mengambil travel insurance

Ever Travelled Abroad

Total : 1692
Total not graduate: 295



Mayoritas orang yang **pernah** keluar negeri **mengambil travel insurance.**

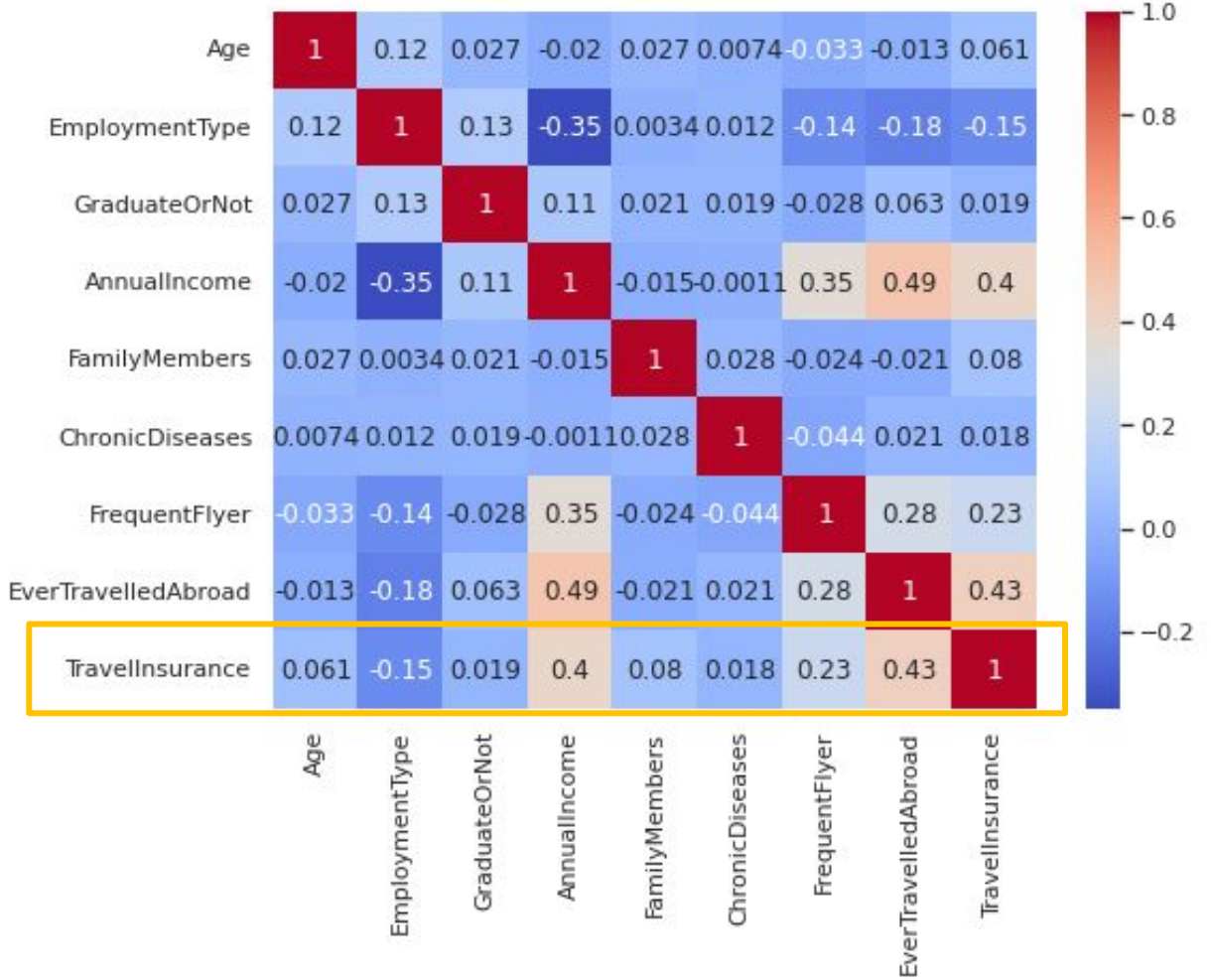
Sedangkan mayoritas (74%) orang yang **belum pernah** keluar negeri **tidak mengambil** travel insurance

Analisis Korelasi
&
Pengecekan Distribusi

Matriks Korelasi

Korelasi terbesar dengan TravelInsurance diperoleh antara variabel **EverTravelAbroad** yaitu sebesar **0.43**

Korelasi variabel **Age, GraduateOrNot, FamilyMembers, dan ChronicDiseases** dengan TravelInsurance sangat rendah (nilainya mendekati nol), sehingga tidak digunakan dalam proses klasifikasi selanjutnya



Pengecekan Distribusi

Jenis Tes: Henze-Zirkler Multivariate Normality Test

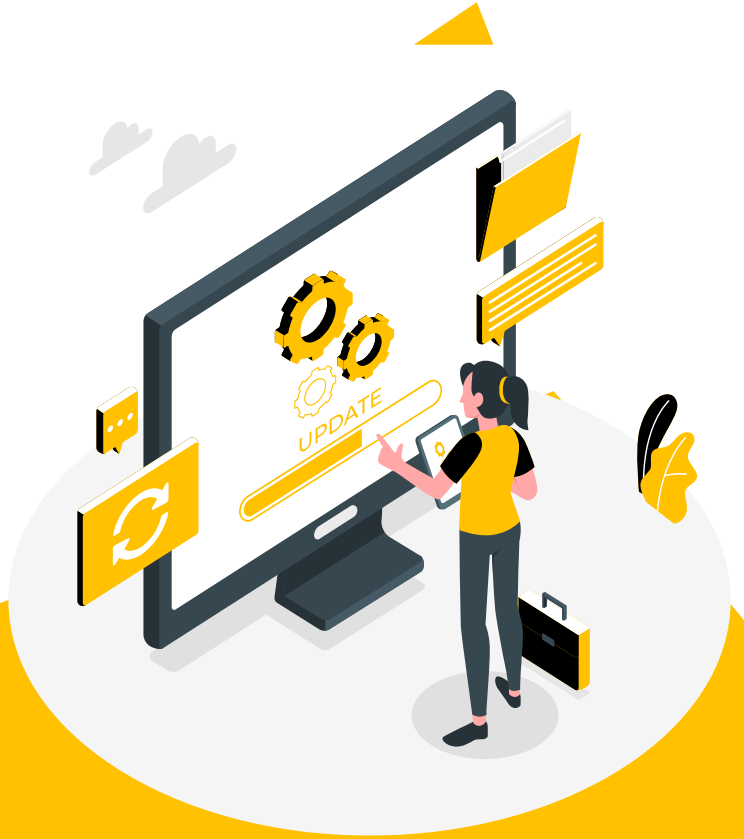
Hz = 334.37

P-Value = 0.0

Normal = False

Data tidak berdistribusi normal. Oleh karena itu dipilih algoritma klasifikasi k-NN dan Decision Tree yang robust terhadap data tidak normal.

Klasifikasi



Klasifikasi Pembelian Travel Insurance

Tujuan Klasifikasi

Menentukan **algoritma klasifikasi terbaik** di antara k-NN dan Decision Tree untuk mengelompokkan pelanggan ke dalam 2 kategori:

1. Membeli travel insurance
2. Tidak membeli travel insurance

Variabel Klasifikasi

- Employment Type
- AnnualIncome
- FrequentFlyer
- EverTravelledAboard

Algoritma Klasifikasi

01

k-Nearest Neighbor

Data uji diklasifikasikan berdasarkan kelas *k* tetangga terdekatnya

Euclidean Distance

$$d(q, x_i) = \sqrt{\sum_{i=1}^{D} (q - x_i)^2}$$

02

Decision Tree

Diagram alur yang menunjukkan jalur yang jelas menuju keputusan

Gini Index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

Entropy

$$Entropy(S) = -\sum_{j=1}^{k} p_j \log_2 p_j$$

Langkah Kerja

01

Split Data

80% Data Training
20% Data Testing

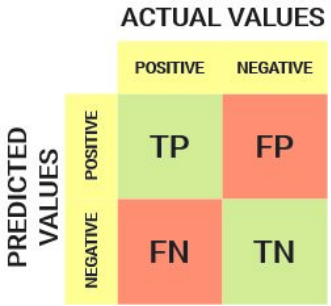
02

Training & Testing

k-Nearest Neighbor
Decision Tree

03

Confusion Matrix



TP: True Positive
FN: False Negative
FP: False Positive
TN: True Negative

04

Menghitung Akurasi, Sensitivitas, dan Spesifisitas

Kinerja Klasifikasi

k-NN

k = 20

Decision Tree

Indeks Gini dan Entropy:

Maximum depth = 5
Random state = 0
Minimum samples leaf = 10
Minimum split = 10

Confusion Matrix

		Actual Values: T	Actual Values: F
Predicted Values	T	63	78
	F	1	256

Akurasi

80.15%

Spesifisitas

100%

Sensitivitas

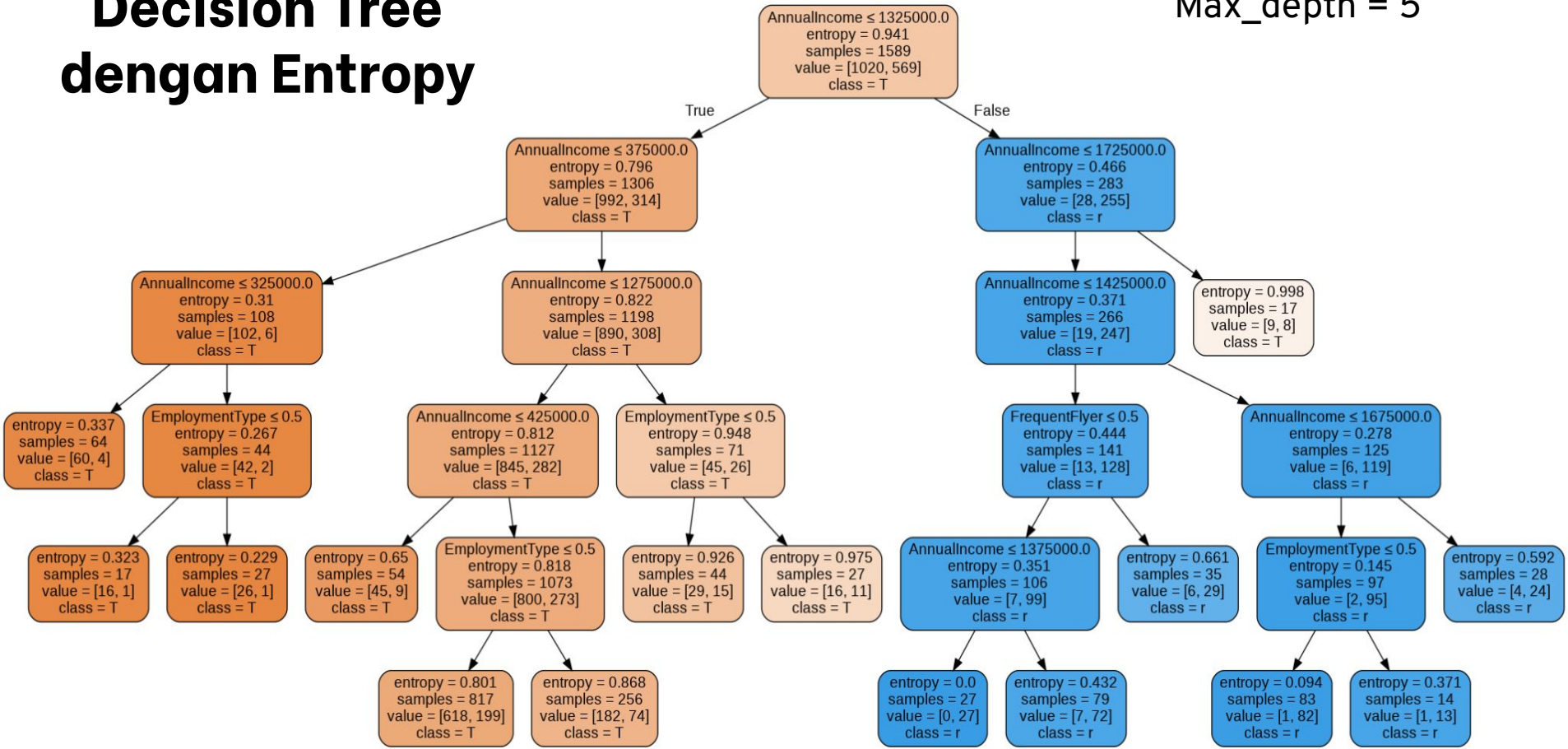
45%

Penyebab: Ketimpangan data

Data TravelInsurance = ‘No’
>
Data TravelInsurance = ‘Yes’

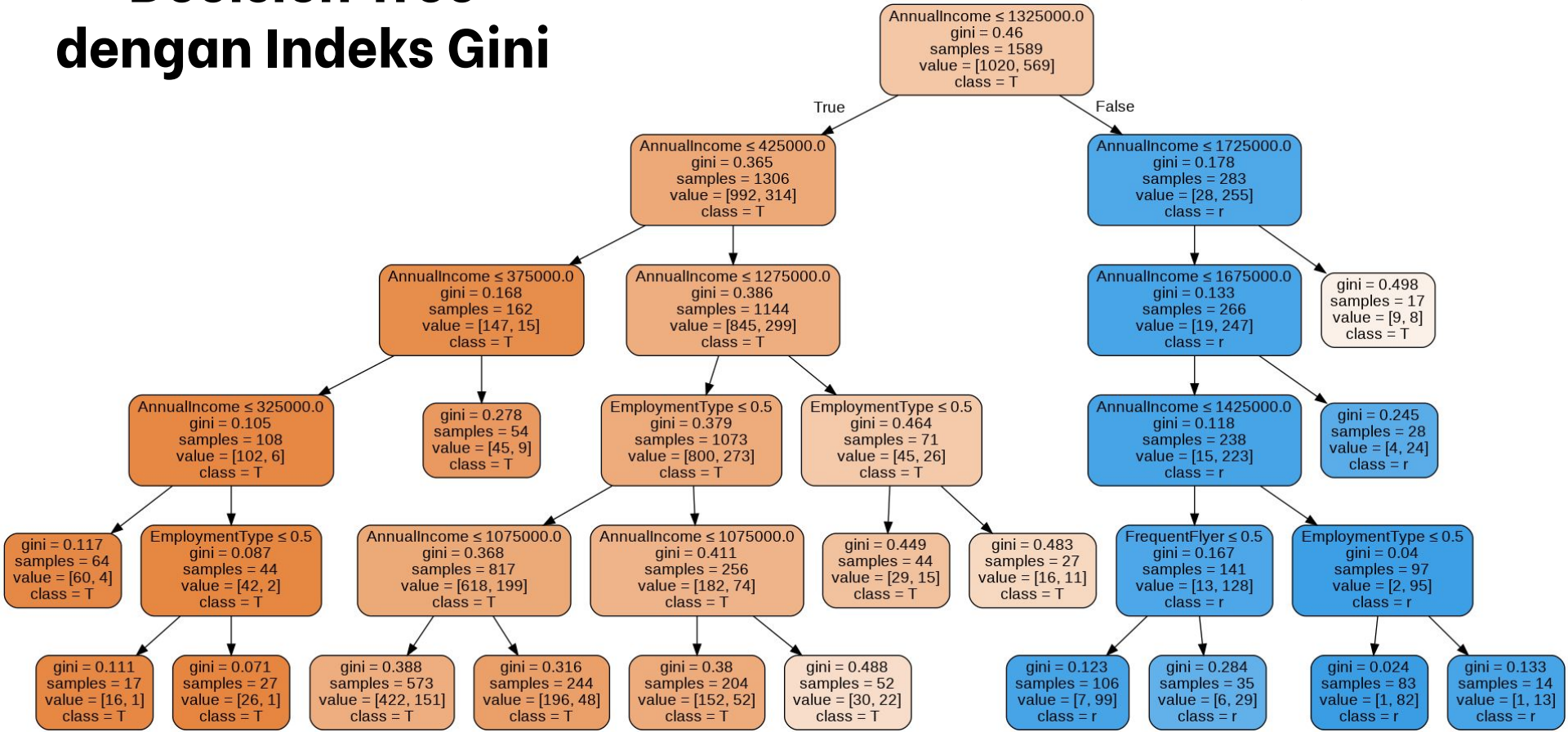
Decision Tree dengan Entropy

Max_depth = 5



Decision Tree dengan Indeks Gini

Max_depth = 5



Kesimpulan

Berdasarkan hasil klasifikasi data TravelInsurancePrediction.csv, diperoleh sensitivitas untuk ketiga algoritma masih tergolong rendah (45%). Hal ini dapat disebabkan oleh ketimpangan dataset pada kelas target Travel Insurance = 1.

Variabel yang memiliki korelasi positif paling kuat dengan Travel Insurance adalah Ever Travel Abroad, yaitu sebesar 0,43 dan Annual Income sebesar 0,4

Nilai akurasi dengan model klasifikasi k-Nearest Neighbor, decision tree indeks gini, dan decision tree indeks entropy memiliki hasil yang sama yaitu sebesar 80,15%. Sehingga ketiga model tersebut baik untuk klasifikasi pada data ini.

Terima Kasih

Lampiran Kode Program

https://colab.research.google.com/drive/1WIx7zwZw1yb7ns3wKO3DVA8N89Z27dTE#scrollTo=nolbKJvtSUAj