



# Penerapan Model Machine Learning untuk Memprediksi *Fraud Auto Insurance*

## Kelompok 4:

Arjun Michael Rogan	2206032860
Audrey Febe Gaberia Siregar	2206052010
Fedora Almanda	2206052004
Sean Felix Fefri Hutagaol	2206051840
Sebastianus Radityo Yoga Pradana	2206051853

# Daftar Isi

Latar Belakang	3
Rumusan Masalah	6
Tujuan Penelitian	7
Alur Simulasi	27
Eksplorasi Data	28
<i>Pre-Processing</i> Data	47
<i>Pipeline</i> Model	60
Optimasi Parameter	61
Evaluasi Kinerja Model	63
Model Terbaik serta <i>Aspek Interpretability</i> dan <i>Explanability</i>	64
Kesimpulan	68

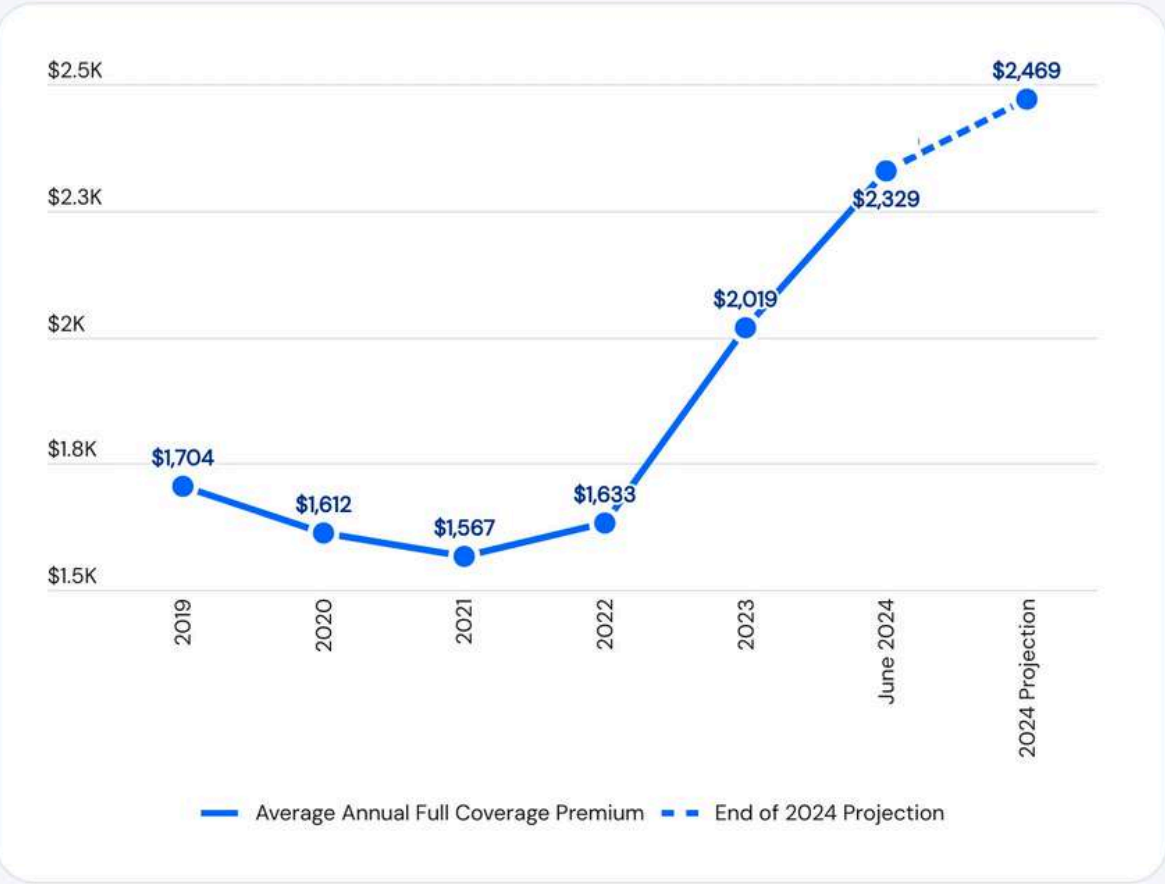


# Latar Belakang

## Auto Insurance?

Auto Insurance merupakan tipe asuransi yang menyediakan perlindungan terhadap kerugian atau kerusakan pada kendaraan seperti mobil, sepeda motor, atau kendaraan komersial, dan lain-lain, tergantung dari kontrak atau *policy* asuransinya

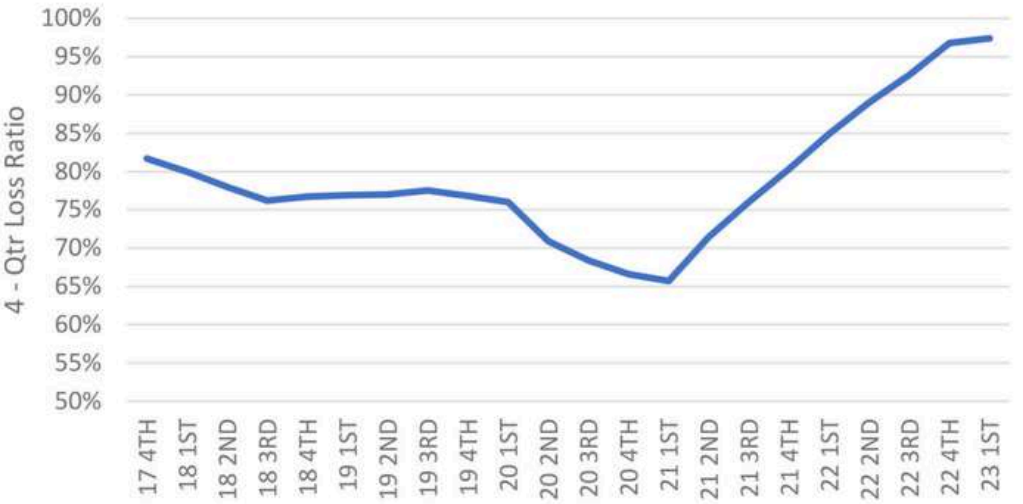
### Average Annual Cost of Full Coverage (2019–2024)



Source: Insurify

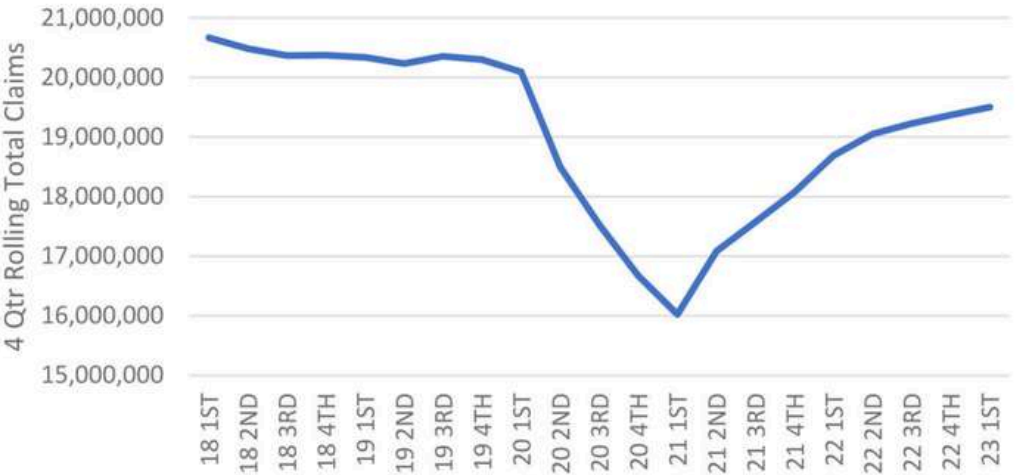
INSURIFY

Personal Auto Rolling 4-Quarter Loss Ratio  
From Fast Track Calendar Quarter Data for All Coverages



Claim Count Trend

From Fast Track Paid Claims for All Coverages 2018 Q1 to 2023 Q1



Source: Fast Track Monitoring System; © Insurance Services Office, Inc., 2023, Chart by Francis Analytics and Actuarial Data Mining.

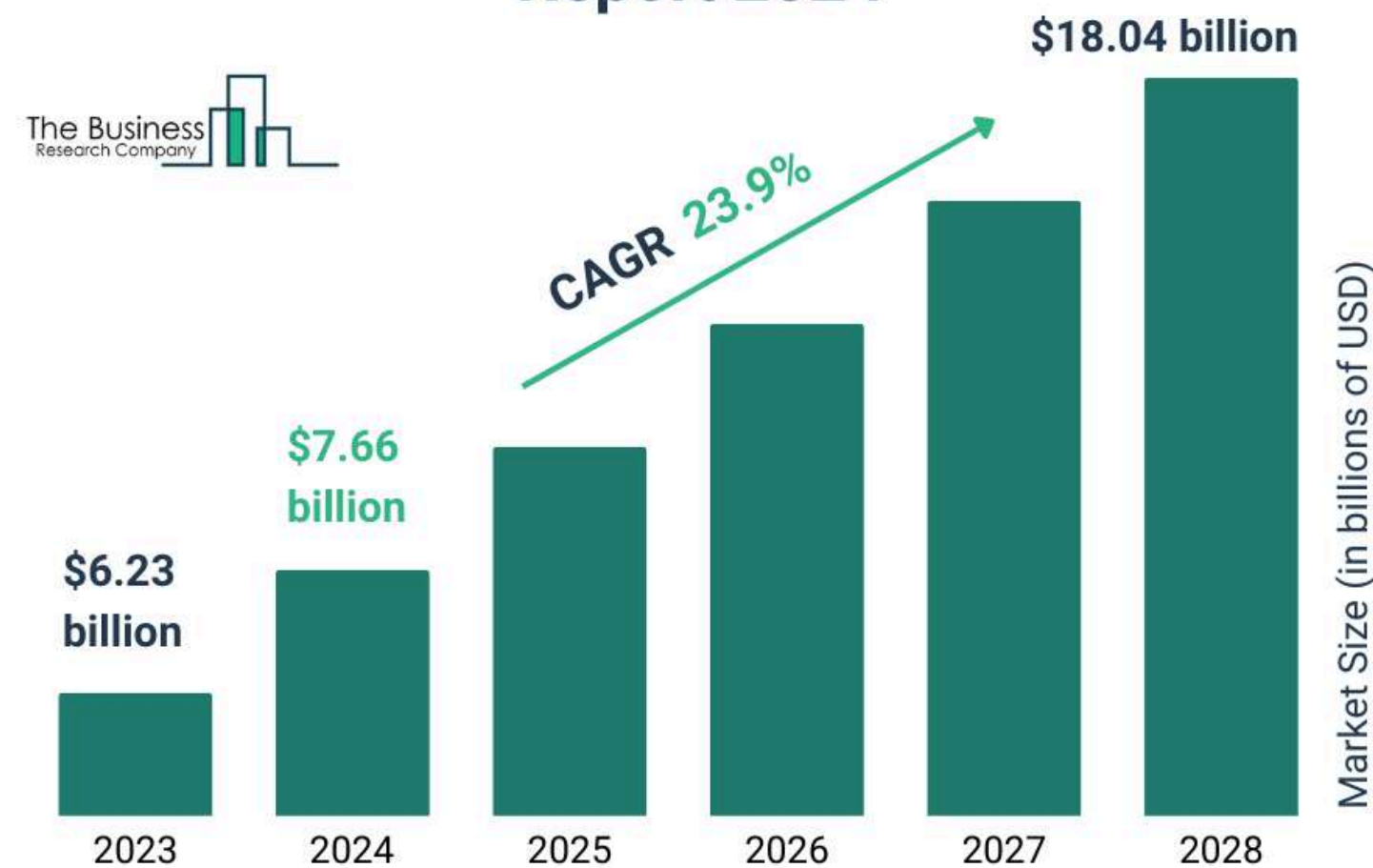
## Fraud On Auto Insurance?

Fraud adalah segala bentuk kecurangan yang dilakukan sepihak demi mendapatkan keuntungan pribadi (AAIA, 2023)

## Dampak Dari Fraud

- Meningkatkan biaya klaim, menurunkan profitabilitas.
- Mendorong kenaikan premi untuk menutup kerugian.
- Memerlukan investasi dalam sistem deteksi dan investigasi fraud.
- Kenaikan premi bagi seluruh pemegang polis, bukan hanya pelaku fraud.
- Mengurangi kepercayaan konsumen terhadap industri asuransi.
- Mengganggu stabilitas ekonomi di sektor asuransi.
- Meningkatkan beban hukum dan pengeluaran terkait pencegahan fraud.
- Mengurangi efektivitas perlindungan konsumen secara keseluruhan.

### Insurance Fraud Detection Global Market Report 2024





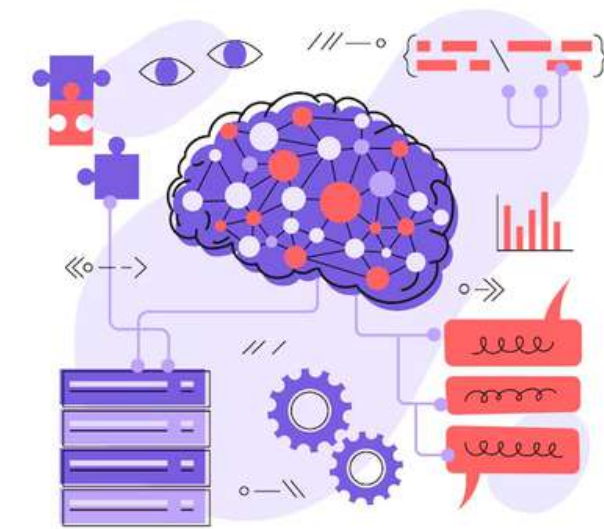
## Machine Learning for Fraud Problems

Untuk mendeteksi *fraud* asuransi, perusahaan asuransi memanfaatkan model *machine learning*.



Raw Data

Data Processing  
& Learning Algorithm



Predictive  
Model

*Machine Learning* dapat membantu memprediksi *fraud* asuransi dengan menganalisis data historis berukuran besar untuk mempelajari pola dan anomali terkait aktivitas penipuan.

# Rumusan Masalah

- 1 Bagaimana perbandingan performa model Logistic Regression, Support Vector Classification (SVC), Random Forest, dan XGBoost?
- 2 Apa bentuk model terbaik untuk memprediksi terjadinya *fraud* pada klaim *auto insurance*?
- 3 Bagaimana aspek *interpretability* dan *explainability* dari model terbaik yang diperoleh dalam mendeteksi *fraud* pada klaim *auto insurance*?



# Tujuan Penelitian

- Membandingkan performa model Logistic Regression, Support Vector Classification (SVC), Random Forest, dan XGBoost.
- Mengetahui model terbaik untuk memprediksi terjadinya fraud pada klaim auto insurance
- Menilai aspek *interpretability* dan *explainability* dari model terbaik



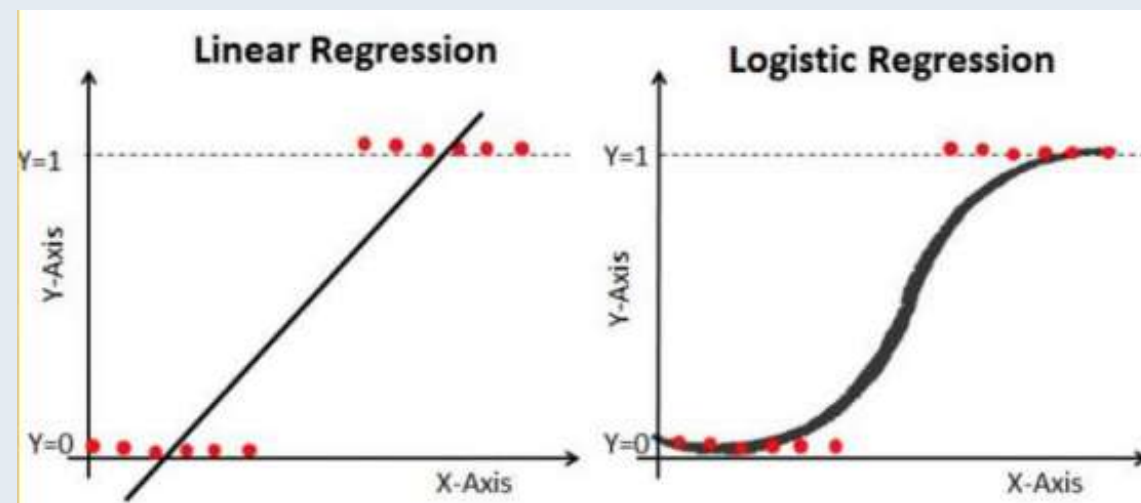
# Landasan Teori

## Logistic Regression

### Definisi Logistic Regression :

Pengembangan regresi linear dengan fungsi logistik untuk data dengan label berupa kelas.

### Visualisasi :



### Rumus - rumus :

Bentuk regresi logistik :

$$p(x) = \sigma(w^T z)$$

$$= \frac{1}{1 + \exp(-(w_0 + w^T x))}$$

dengan

$$z = (1, x_1, \dots, x_D)^T$$

$$w = (w_0, w_1, \dots, w_D)^T$$



# Landasan Teori

## Support Vector Machine (SVM)

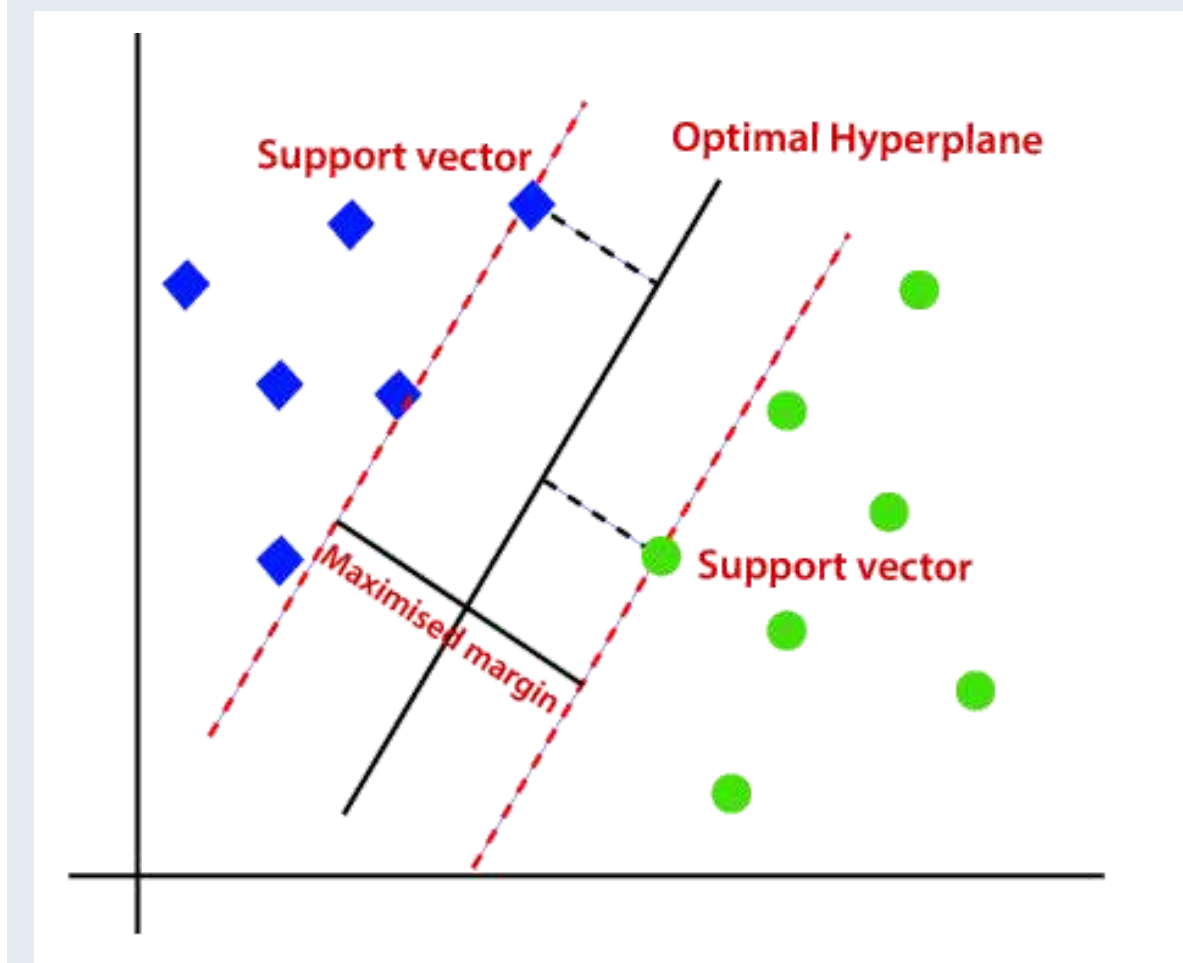
### Definisi SVM

Algoritma pembelajaran mesin yang digunakan untuk klasifikasi dan regresi dengan menemukan hyperplane optimal yang memisahkan data ke dalam kelas-kelas berbeda dengan margin maksimal.

### Fungsi Kernel :

Fungsi Kernel adalah fungsi yang mengukur kemiripan antara dua vektor data tersebut dalam ruang asli, yang setara dengan melakukan produk dalam ruang berdimensi lebih tinggi tanpa melakukan transformasi eksplisit.

### Visualisasi :



# Landasan Teori

## Support Vector Machine (SVM)

### Formula:

**Fungsi Optimasi Margin Lunak:**

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2} \|w\|^2 + C \sum_{n=1}^N \xi_n \\ \text{s.t.} \quad & t_n (w^T \phi(x_n) + b) \geq 1 - \xi_n, \quad n = 1, \dots, N \\ & \xi_n \geq 0 \end{aligned}$$

**Fungsi Dual Margin Lunak:**

$$\begin{aligned} \max_a \quad & \tilde{L}(a) = \sum_{n=1}^N a_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N a_n a_m t_n t_m k(x_n, x_m) \\ \text{s.t.} \quad & 0 \leq a_n \leq C, \quad n = 1, \dots, N \\ & \sum_{n=1}^N a_n t_n = 0 \end{aligned}$$

**Fungsi Prediksi :**

$$\max_{w,b} \left\{ \frac{1}{\|w\|} \min_n [t_n (w^T \phi(x_n) + b)] \right\}$$

**Fungsi Prediksi :**

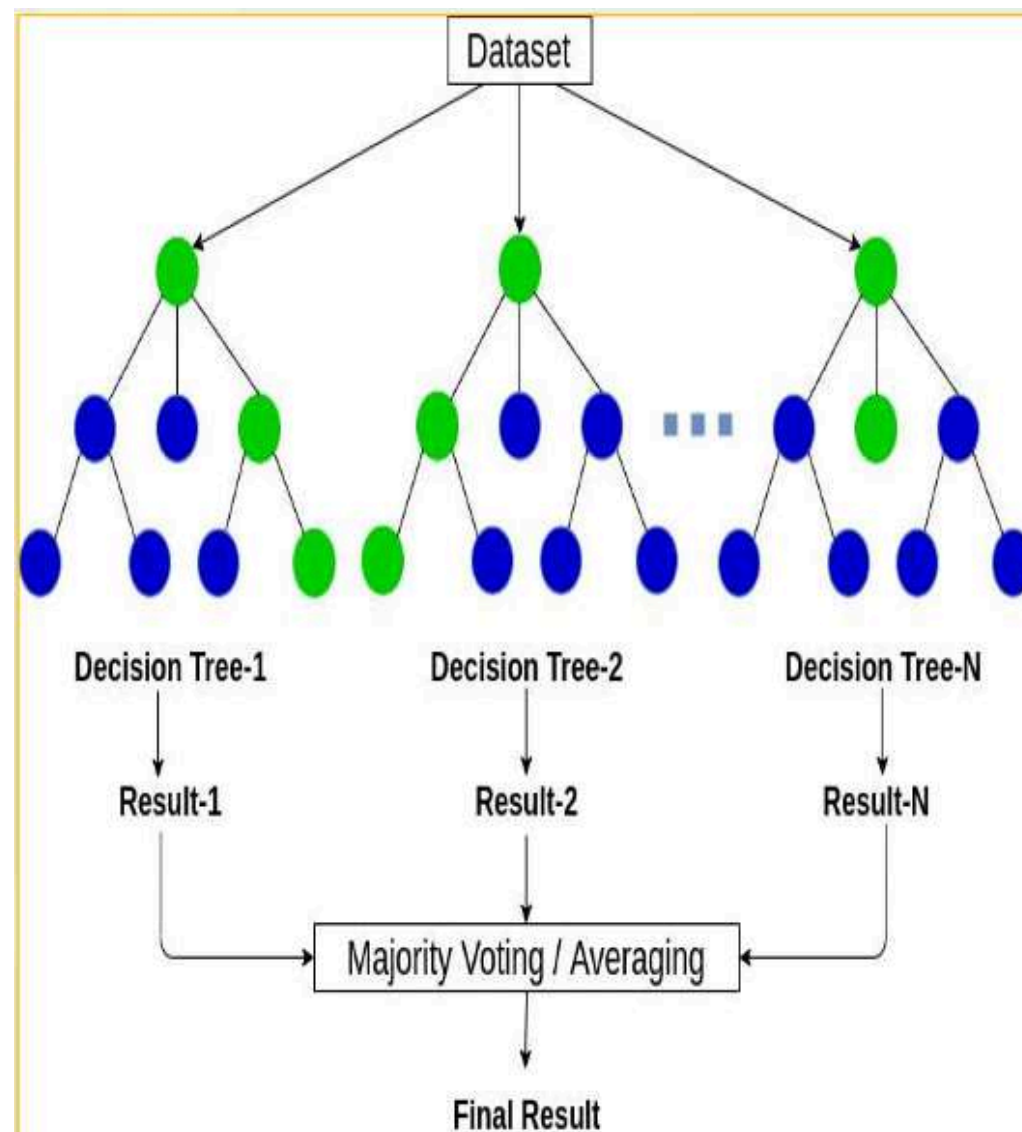
$$f(\mathbf{x}) = d(y(x)) = d \left( \sum_{m \in 1} a_m y_m K(\mathbf{x}, \mathbf{x}_m) + b \right)$$



# Landasan Teori

## Random Forest

### ILUSTRASI



### Definisi Random Forest

Random Forest merupakan model *machine learning* yang mengimplementasikan *ensemble learning*, yaitu *bagging*, terhadap model Decision Tree. Random Forest mampu menangani masalah overfitting pada model Decision Tree.

### *Bagging*

*Bagging* merupakan pendekatan pada *ensemble learning* di manan beberapa model dasar dibangun secara independen dan prediksi akhir adalah pemilihan (voting) dari masing-masing model dasar tersebut.

# Landasan Teori

## Random Forest

### Rumus - rumus :

#### Gini Impurity

$$G = 1 - \sum_{i=1}^k p_i^2$$

#### Recursive Partitioning

$$\text{Best Split} = \operatorname{argmin} [w_{\text{left}} I_{\text{left}} + w_{\text{right}} I_{\text{right}}]$$

#### Bagging

$$D_b = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$$

#### Out-of-Bag Error

$$OOB = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(y_i \neq \hat{y}_{OOB,i})$$

#### Ensemble Prediction

$$\hat{y} = \operatorname{mode}(\{T_1(x), T_2(x), \dots, T_k(x)\})$$



# Landasan Teori

## eXtreme Gradient Boosting (XGBoost)

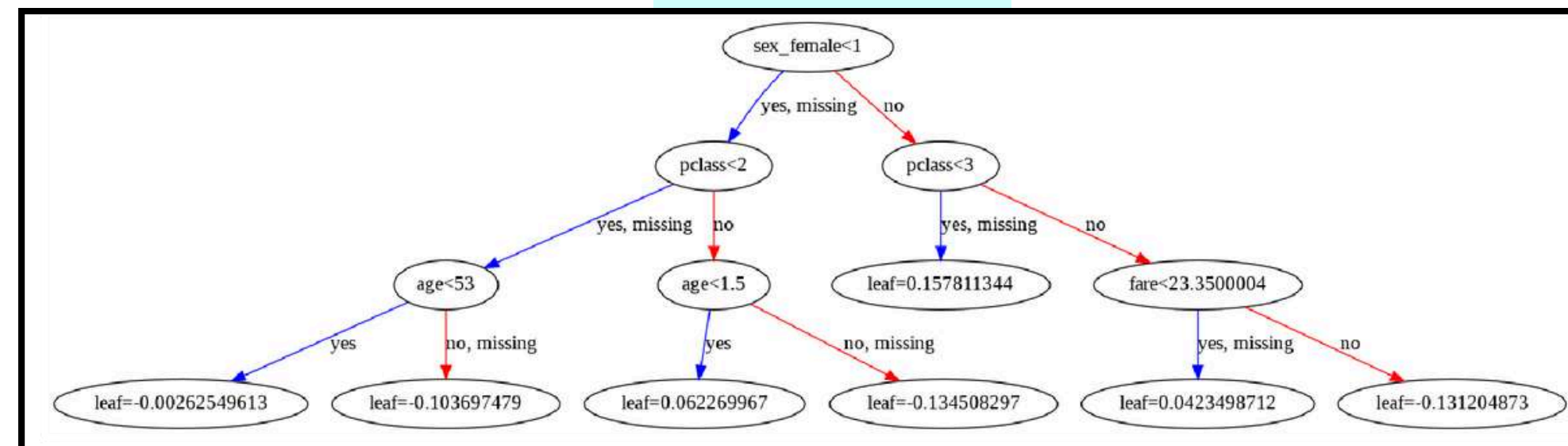
### Definisi XGBoost :

Implementasi yang dioptimalkan dari algoritma Gradient Boosting.

### Gradient Boosting :

Teknik ensemble dalam machine learning untuk membangun model prediksi yang kuat dari kombinasi beberapa model yang lemah.

### ILUSTRASI



Sumber: [https://mikulskibartosz.name/images/2019-08-26-how-to-plot-the-decision-trees-from-xgboost-classifier/xgboost\\_tree.pngtext](https://mikulskibartosz.name/images/2019-08-26-how-to-plot-the-decision-trees-from-xgboost-classifier/xgboost_tree.pngtext)

# Landasan Teori

## eXtreme Gradient Boosting (XGBoost)

### Rumus - rumus :

#### Objective Function

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

#### Split Gain

$$\text{Gain} = \frac{1}{2} \left[ \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \frac{G_L^2}{H_L + \lambda} - \frac{G_R^2}{H_R + \lambda} \right] - \gamma$$

#### Leaf Weight

$$w_j = - \frac{\sum_{i \in \text{leaf}_j} g_i}{\sum_{i \in \text{leaf}_j} h_i + \lambda}$$

#### Learning Rate

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$



# Landasan Teori

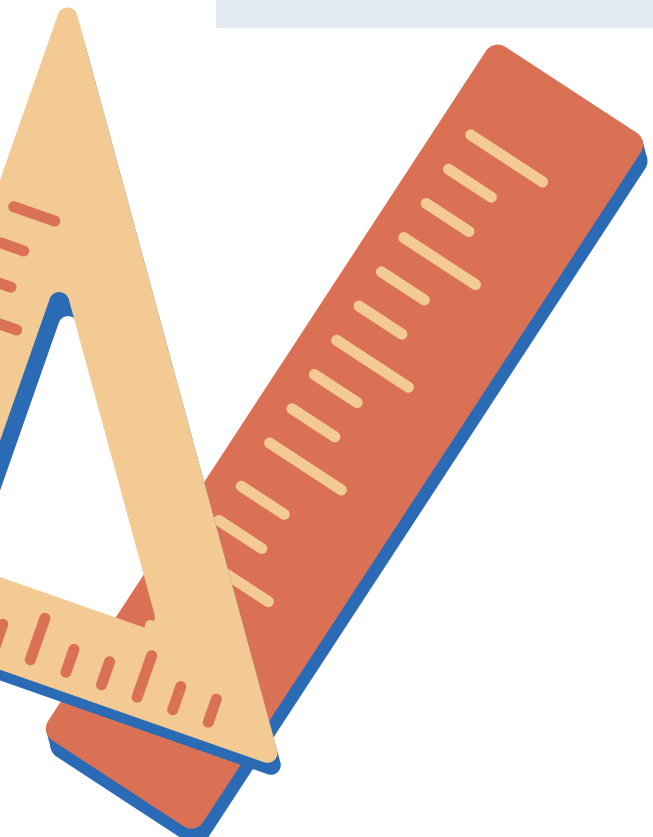
## Min-Max Scaling

### Definisi Scaling

Proses penyesuaian rentang nilai dan distribusi dari fitur-fitur yang dipakai dalam Machine Learning.

### Fungsi Scaling

Tujuan scaling adalah untuk memastikan bahwa tidak ada fitur yang mendominasi fitur lain. Dominasi suatu fitur bisa disebabkan oleh perbedaan rentang nilai, yang mana dapat mempengaruhi kinerja suatu machine learning.



### Min-Max Scaling

Proses ini mengskalakan setiap fitur numerik ke dalam range tetap [0,1]. Min-Max Scaling dilakukan ketika dataset memiliki rentang nilai yang ekstrem dan tidak berdistribusi normal.

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Keterangan:

$X'$  : Nilai setelah scaling

$X_{min}$  : Nilai minimum fitur

$X$  : Nilai awal

$X_{max}$  : Nilai maksimum fitur

# Landasan Teori

## Synthetic Minority Oversampling Technique (SMOTE)

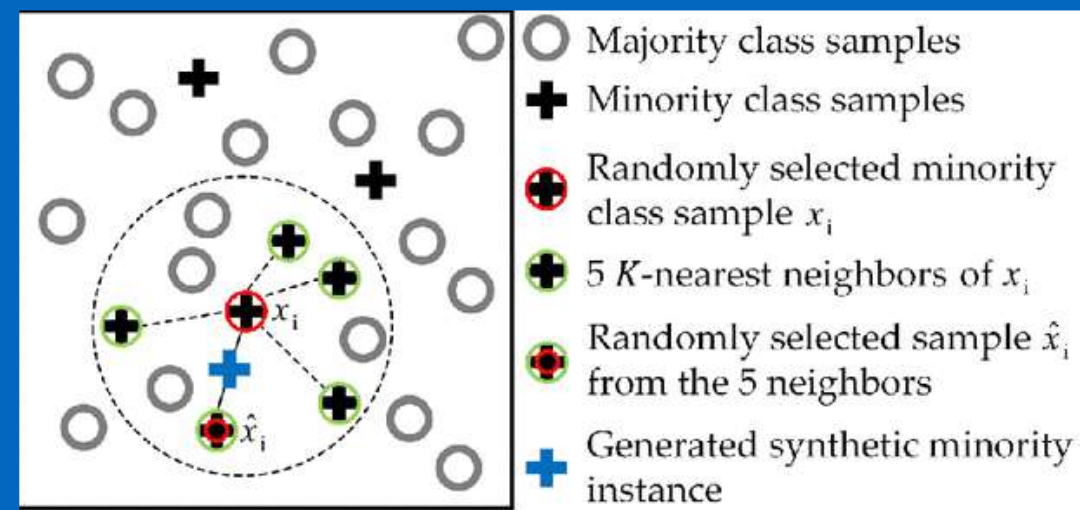
### Definisi

Teknik yang digunakan untuk menangani masalah ketidakseimbangan kelas dalam dataset

### Kegunaan

Untuk mengurangi bias dalam model *machine learning* yang disebabkan oleh ketidakseimbangan jumlah data antara kelas-kelas target.

### Visualisasi SMOTE



Sumber: [https://rikunert.com/smote\\_explained](https://rikunert.com/smote_explained)

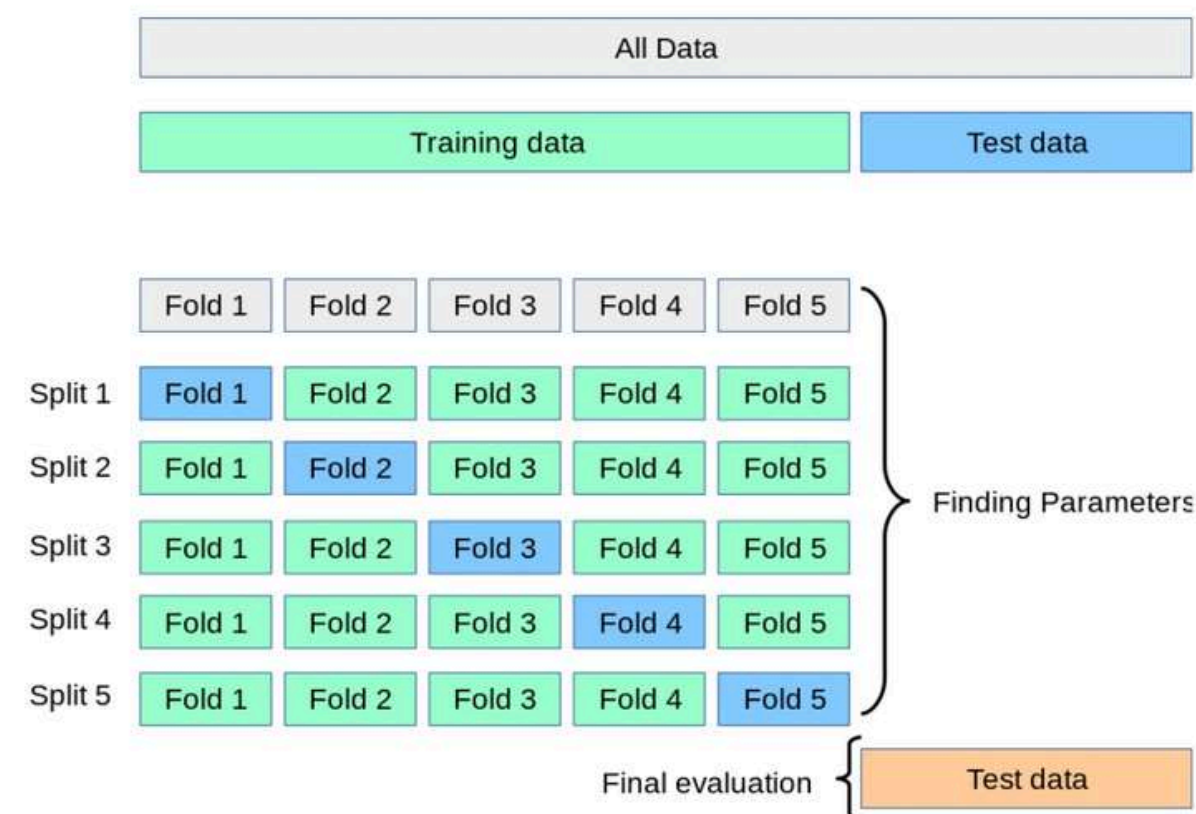


# Landasan Teori

## Hyperparameter Tuning

### Definisi

Hyperparameter Tuning adalah proses optimalisasi *hyperparameter* model *machine learning* untuk meningkatkan akurasi model dan mencegah terjadinya *overfitting* dan *underfitting*.



Sumber: [https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRPDzu1EInJLRQoCrOR-fOeBBI\\_mhr92zaLpg&s](https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRPDzu1EInJLRQoCrOR-fOeBBI_mhr92zaLpg&s)

### Grid Search and Cross Validation

- Metode *hyperparameter tuning* di mana algoritmanya menguji berbagai kombinasi *hyperparameter* untuk menemukan kombinasi *hyperparameter* terbaik.
- *Cross-validation* dilakukan pada setiap kombinasi *hyperparameter* untuk mencegah *overfitting*.
- Data dilatih dan diuji pada subset data yang berbeda-beda.

# Landasan Teori

## Area Under the Curve (AUC) - Receiver Operating Characteristic (ROC)

### Definisi

Metrik evaluasi yang digunakan untuk mengukur kinerja model klasifikasi.

### AUC

Luas area di bawah kurva ROC yang memberikan nilai seberapa baik model dapat membedakan antara kelas positif dan negatif.

Nilai AUC berkisar antara 0 hingga 1, dengan interpretasi sebagai berikut:

- **AUC = 1** (Model sempurna dalam membedakan kelas.)
- **AUC = 0.5** (Model tidak memiliki kemampuan memisahkan kelas.)
- **AUC < 0.5** (Model berkinerja lebih buruk daripada prediksi acak.)

Reference : Çorbacioğlu ŞK, Aksel G. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. Turk J Emerg Med. 2023 Oct 3;23(4):195-198. doi: 10.4103/tjem.tjem\_182\_23. PMID: 38024184; PMCID: PMC10664195.

# Landasan Teori

## Area Under the Curve (AUC) - Receiver Operating Characteristic (ROC)

### ROC

Kurva yang menampilkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) untuk berbagai threshold keputusan yang digunakan oleh model.

### TPR/sensitivity/recall

- ▶ Proporsi data positif yang diprediksi benar oleh model.

$$TPR = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

### FPR

- ▶ Proporsi data negatif yang salah diprediksi sebagai positif oleh model.

$$FPR = \frac{\text{False Positive (FP)}}{\text{False Positive (FP)} + \text{True Negative (TN)}}$$



# Landasan Teori

## Area Under the Curve (AUC) - Receiver Operating Characteristic (ROC)

### Interpretasi AUC - ROC

<b>0.5 - 0.6</b>	<b>Failed</b>
<b>0.6 - 0.7</b>	<b>Worthless</b>
<b>0.7 - 0.8</b>	<b>Poor</b>
<b>0.8 - 0.9</b>	<b>Good</b>
<b>&gt; 0.9</b>	<b>Excellent</b>

sumber:

Polo TCF, Miot HA. Use of ROC curves in clinical and experimental studies. J Vasc Bras. 2020;19: e20200186.

<https://doi.org/10.1590/1677-5449.200186>

# Landasan Teori

## Confusion Matrix

### Confusion Matrix

Tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan cara menunjukkan hasil prediksi model terhadap data uji dibandingkan dengan nilai sebenarnya.

	Prediksi Negatif	Prediksi Positif
Aktual Negatif	True Negative (TN)	False Positive (FP)
Aktual Positif	False Negative (FN)	True Positive (TP)

### Metrik evaluasi yang dapat dihitung :

- Accuracy : Mengukur proporsi prediksi benar dari keseluruhan prediksi.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precision : Mengukur proporsi prediksi positif yang benar dari semua prediksi positif.

$$Precision = \frac{TP}{TP + FP}$$

- Recall : menunjukkan seberapa baik model dalam menangkap semua kasus positif

$$Recall = \frac{TP}{TP + FN}$$

# Landasan Teori

## Confusion Matrix

	Prediksi Negatif	Prediksi Positif
Aktual Negatif	True Negative (TN)	False Positive (FP)
Aktual Positif	False Negative (FN)	True Positive (TP)

### Metrik evaluasi yang dapat dihitung :

- F-1 Score : rata-rata harmonis dari precision dan recall.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- False Positive Rate (FPR) : Mengukur proporsi prediksi positif yang salah dari semua kasus negatif yang ada.

$$FPR = \frac{FP}{FP + TN}$$

- Specificity (True Negative Rate) : Mengukur proporsi prediksi negatif yang benar dari semua kasus negatif yang ada.

$$Specificity = \frac{TN}{TN + FP}$$



# Landasan Teori

## Feature Importance

### Definisi

Ukuran kontribusi setiap fitur terhadap prediksi yang dihasilkan

Ukuran dapat dihitung berdasarkan metrik **Gain**:

- Mengukur rata-rata penurunan loss
- Peningkatan gain yang signifikan --> fitur penting

### Rumus-rumus Feature Importance

$$\text{Gain}_j = \frac{1}{N_j} \sum_{\text{split}} \text{Gain}_{\text{split}} \quad N_j : \text{jumlah total split}, \text{Gain}_{\text{split}} : \text{peningkatan akurasi}$$

$$\text{Gain}_{\text{split}} = \frac{1}{2} \left( \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right) - \gamma$$

$G_L$  dan  $H_L$ : Total gradien dan hessian di cabang kiri setelah split.

$G_R$  dan  $H_R$ : Total gradien dan hessian di cabang kanan setelah split.

$\lambda$ : Parameter regularisasi L2.

$\gamma$ : Penalti pruning.

# Landasan Teori

## Partial Dependence Plot (PDP)

### Definisi

Grafik yang menunjukkan ketergantungan antara respons dan satu atau beberapa fitur masukan, dengan mengabaikan nilai semua fitur masukan lainnya.

### Jenis Ilustrasi :

#### One-Way

One-Way PDP menunjukkan hubungan antara satu fitur dengan prediksi model dan fitur lain dianggap tetap. Biasanya divisualisasikan dengan grafik garis.

#### Two-Way

Two-Way PDP memperlihatkan interaksi antara dua fitur dengan prediksi model. PDP ini biasanya divisualisasikan dalam bentuk peta kontur (heatmap) atau permukaan 3D.

# Landasan Teori

## Partial Dependence Plot (PDP)

### Rumus

Untuk satu fitur  $X(j)$ , PDP didefinisikan sebagai:

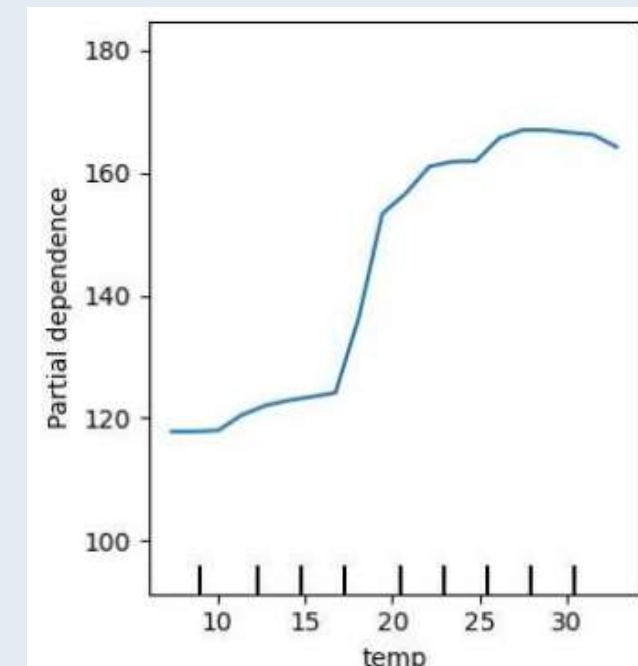
$$PD(X_j) = \frac{1}{n} \sum_{i=1}^n \hat{f}(X_j, X_{-j}^{(i)})$$

Keterangan :

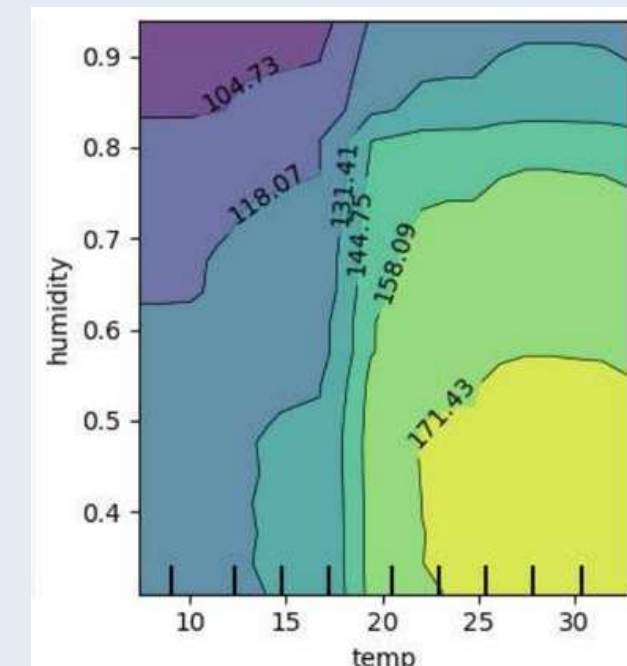
- $\hat{f}$ : fungsi prediksi model
- $X_j$ : fitur target yang ingin dipelajari
- $X_{-j}^{(i)}$ : nilai dari semua fitur lain kecuali  $X_j$  untuk data ke- $i$
- $n$ : jumlah sampel data

### Contoh Plot

One-Way



Two-Way



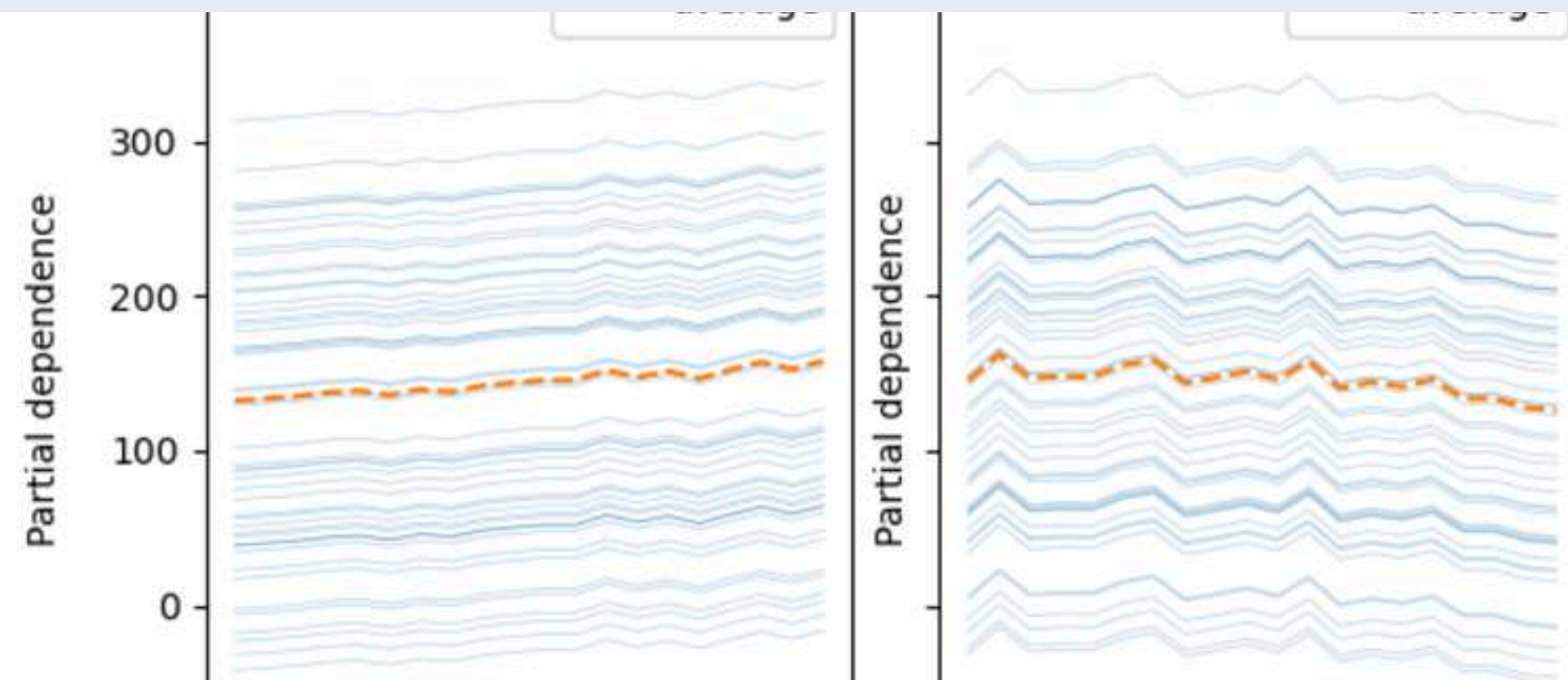


# Landasan Teori

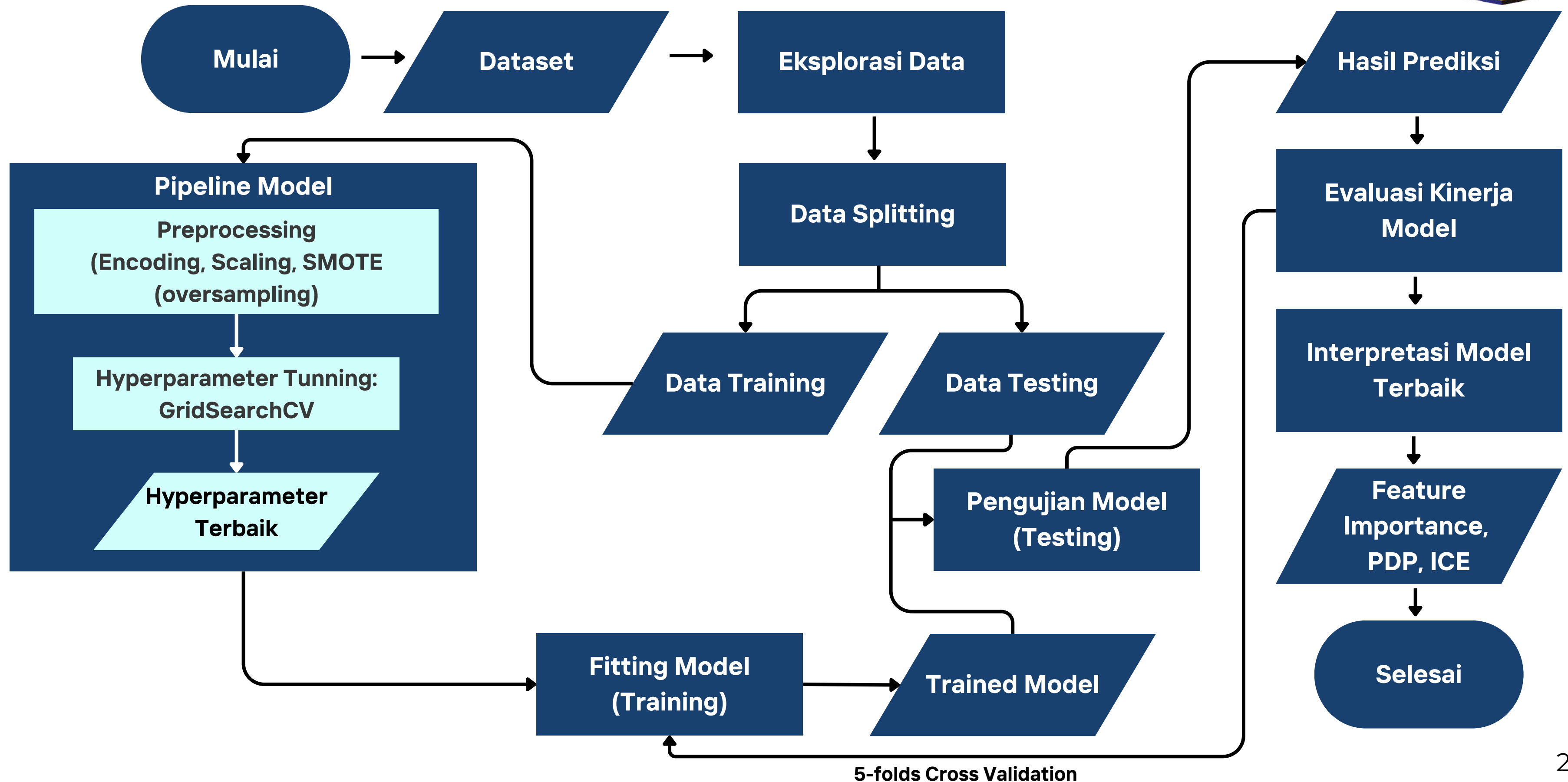
## Individual Conditional Expectation (ICE)

### Definisi

Individual Conditional Expectation (ICE) adalah teknik yang digunakan dalam analisis model prediktif untuk memahami bagaimana prediksi model berubah seiring perubahan pada satu variabel input, dengan mempertahankan nilai variabel lainnya tetap konstan. Berbeda dengan Partial Dependence Plots (PDP), yang menggambarkan hubungan rata-rata antara fitur dan target, ICE menggambarkan prediksi untuk setiap individu dalam dataset untuk berbagai nilai fitur.



# Alur Simulasi



# Eksplorasi Data

Identifikasi  
Fitur Numerik

- |   |  |
|---|--|
| 1 | months_as_customer: jumlah bulan terdaftar asuransi sebelum polis autoinsurance                                  |
| 2 | age: usia pemegang polis.  |
| 3 | policy_number: nomor indentifikasi unik pemegang polis.  |
| 4 | policy_deductable: deductible setiap klaim.  |
| 5 | policy_annual_premium: jumlah premi tahunan.   |
| 6 | umbrella_limit: Coverage limit under an umbrella insurance policy, which provides additional liability coverage. |
| 7 | capital_gains: Capital gains reported by the insured, relevant to their financial profile.                       |
| 8 | capital_loss: Capital losses reported by the insured, which also relate to financial circumstances.              |
| 9 | incident_hour_of_the_day: jam kejadian kecelakaan  |



# Eksplorasi Data

Identifikasi  
Fitur Numerik

10	number_of_vehicles_involved: Jumlah kendaraan yang terlibat dalam kecelakaan
11	bodily_injuries: Jumlah orang yang mengalami cedera pada kecelakaan
12	witnesses: Jumlah saksi yang melaporkan kecelakaan
13	total_claim_amount: Jumlah total klaim dalam dolar.
14	injury_claim: Bagian dari jumlah total klaim yang disebabkan oleh cedera.
15	property_claim: Bagian dari jumlah total klaim yang disebabkan oleh kerusakan properti.
16	vehicle_claim: Bagian dari jumlah total klaim yang disebabkan oleh kerusakan kendaraan
17	auto_year: Tahun produksi dari kendaraan pihak bertanggung.

# Eksplorasi Data

Identifikasi  
Fitur Kategorik  
Nominal

- |    |   |
|----|---|
| 1  | policy_bind_date: tanggal pemegang polis bergabung ke polis asuransi                      |
| 2  | policy_state: negara bagian polis asuransi didaftarkan                                    |
| 3  | insured_zip: kode pos pemegang polis asuransi   |
| 4  | insured_sex: gender pemegang polis (M untuk laki-laki dan F untuk perempuan)              |
| 5  | insured_occupation: pekerjaan pemegang polis  |
| 6  | insured_hobbies: hobi pemegang polis  |
| 7  | insured_relationship: status hubungan dan keluarga pemegang polis                         |
| 8  | incident_date: tanggal kecelakaan   |
| 9  | incident_type: tipe kecelakaan  |
| 10 | collision_type: area terdampak pada kecelakaan (contoh: tabrakan depan, tabrakan samping) |

# Eksplorasi Data

Identifikasi  
Fitur Kategorik  
Nominal

- |    |   |
|----|---|
| 11 | authorities_contacted: Pihak berwenang yang dikontak setelah terjadi kecelakaan (polisi, pemadam kendaraan, ambulans, lainnya, tidak ada) |
| 12 | incident_state: Negara bagian di mana kecelakaan terjadi.   |
| 13 | incident_city: Kota di mana kecelakaan terjadi  |
| 14 | incident_location: Alamat atau lokasi spesifik terjadinya kecelakaan.   |
| 15 | property_damage: Indikator terjadinya kerusakan properti (Ya/Tidak)   |
| 16 | police_report_available Indikator ketersediaan laporan polisi (Ya/Tidak)  |
| 17 | auto_make: Merk dari kendaraan pihak tertanggung.   |
| 18 | auto_model: Model dari kendaraan pihak tertanggung.   |
| 19 | fraud_reported: Indikator klaim dilaporkan sebagai fraud (Ya/Tidak)   |
| 20 | policy_csl: batas besar klaim (per orang/per insiden).  |

# Eksplorasi Data

Identifikasi  
Fitur Kategorik  
Ordinal

- 
- |   |   |
|---|---|
| 1 | incident_severity: tingkat keparahan kecelakaan (contoh: minor, major, total loss). |
| 2 | insured_education_level: tingkat pendidikan pemegang polis                          |
-



# Eksplorasi Data

## Cuplikan 5 entri pertama

	months_as_customer	age	policy_number	policy_bind_date	policy_state	policy_csl	policy_deductable	policy_annual_premium	umbrella_limit
0	60	23	842643	1997-11-20	OH	500/1000	500	1215.36	3000000
1	370	55	285496	1994-05-27	IL	100/300	2000	972.18	0
2	413	55	115399	1991-02-08	IN	100/300	2000	1268.79	0
3	64	25	908616	2000-02-18	IL	250/500	1000	954.16	0
4	114	30	584859	1992-04-04	IL	100/300	1000	1558.29	0

insured_zip	insured_sex	insured_education_level	insured_occupation	insured_hobbies	insured_relationship	capital-gains	capital-loss	incident_date
432220	MALE	MD	protective-serv	reading	wife	0	0	2015-01-22
443920	MALE	High School	prof-specialty	paintball	other-relative	72700	-68200	2015-01-11
453148	MALE	MD	priv-house-serv	chess	own-child	0	-31000	2015-01-19
473328	MALE	Masters	prof-specialty	video-games	husband	53200	0	2015-01-18
472248	MALE	High School	farming-fishing	video-games	wife	51400	-64000	2015-01-09

incident_type	collision_type	incident_severity	authorities_contacted	incident_state	incident_city	incident_location	incident_hour_of_the_day
Single Vehicle Collision	Rear Collision	Total Loss	Ambulance	SC	Northbend	6655 5th Drive	9
Multi-vehicle Collision	Rear Collision	Major Damage	Ambulance	SC	Hillsdale	2526 Embaracadero Ave	20
Single Vehicle Collision	Front Collision	Total Loss	Ambulance	WV	Northbend	5667 4th Drive	15
Multi-vehicle Collision	Side Collision	Major Damage	Ambulance	SC	Columbus	4687 5th Drive	22
Multi-vehicle Collision	Front Collision	Major Damage	Ambulance	NY	Hillsdale	8353 Britain Ridge	1

number_of_vehicles_involved	property_damage	bodily_injuries	witnesses	police_report_available	total_claim_amount	injury_claim	property_claim
1	YES	1	0	NO	56520	4710	9420
3	NO	0	0	YES	71520	17880	5960
1	?	2	2	?	98160	8180	16360
4	NO	0	0	?	75600	12600	12600
3	NO	1	2	?	77110	14020	14020

vehicle_claim	auto_make	auto_model	auto_year	fraud_reported	_c39
42390	Saab	95	2000	N	NaN
47680	Suburu	Forrestor	2000	Y	NaN
73620	Dodge	RAM	2011	Y	NaN
50400	Toyota	Corolla	2005	N	NaN
49070	Suburu	Impreza	2015	N	NaN

1000 entri data; 39 variabel



# Eksplorasi Data

## Summary Data

	months_as_customer	age	policy_number \		policy_bind_date	policy_deductable \
count	1000.000000	1000.000000	1000.000000	count	1000	1000.000000
mean	203.954000	38.948000	546238.648000	mean	2002-02-08 04:40:47.999999872	1136.000000
min	0.000000	19.000000	100804.000000	min	1990-01-08 00:00:00	500.000000
25%	115.750000	32.000000	335980.250000	25%	1995-09-19 00:00:00	500.000000
50%	199.500000	38.000000	533135.000000	50%	2002-04-01 12:00:00	1000.000000
75%	276.250000	44.000000	759099.750000	75%	2008-04-21 12:00:00	2000.000000
max	479.000000	64.000000	999435.000000	max	2015-02-22 00:00:00	2000.000000
std	115.113174	9.140287	257063.005276	std	NaN	611.864673

	policy_annual_premium	umbrella_limit	insured_zip	capital-gains \
count	1000.000000	1.000000e+03	1000.000000	1000.000000
mean	1256.406150	1.101000e+06	501214.488000	25126.100000
min	433.330000	-1.000000e+06	430104.000000	0.000000
25%	1089.607500	0.000000e+00	448404.500000	0.000000
50%	1257.200000	0.000000e+00	466445.500000	0.000000
75%	1415.695000	0.000000e+00	603251.000000	51025.000000
max	2047.590000	1.000000e+07	620962.000000	100500.000000
std	244.167395	2.297407e+06	71701.610941	27872.187708



# Eksplorasi Data

## Summary Data

	capital-loss	...	incident_hour_of_the_day	\		number_of_vehicles_involved	bodily_injuries	witnesses	\
count	1000.000000	...	1000.000000		count	1000.000000	1000.000000	1000.000000	
mean	-26793.700000	...	11.644000		mean	1.83900	0.992000	1.487000	
min	-111100.000000	...	0.000000		min	1.00000	0.000000	0.000000	
25%	-51500.000000	...	6.000000		25%	1.00000	0.000000	1.000000	
50%	-23250.000000	...	12.000000		50%	1.00000	1.000000	1.000000	
75%	0.000000	...	17.000000		75%	3.00000	2.000000	2.000000	
max	0.000000	...	23.000000		max	4.00000	2.000000	3.000000	
std	28104.096686	...	6.951373		std	1.01888	0.820127	1.111335	

	total_claim_amount	injury_claim	property_claim	vehicle_claim	\		auto_year	_c39
count	1000.00000	1000.000000	1000.000000	1000.000000		count	1000.000000	0.0
mean	52761.94000	7433.420000	7399.570000	37928.950000		mean	2005.103000	NaN
min	100.00000	0.000000	0.000000	70.000000		min	1995.000000	NaN
25%	41812.50000	4295.000000	4445.000000	30292.500000		25%	2000.000000	NaN
50%	58055.00000	6775.000000	6750.000000	42100.000000		50%	2005.000000	NaN
75%	70592.50000	11305.000000	10885.000000	50822.500000		75%	2010.000000	NaN
max	114920.00000	21450.000000	23670.000000	79560.000000		max	2015.000000	NaN
std	26401.53319	4880.951853	4824.726179	18886.252893		std	6.015861	NaN

# Eksplorasi Data

## Tipe Data dari Seluruh Fitur

Data columns (total 40 columns):

#	Column	Non-Null Count	Dtype
0	months_as_customer	1000 non-null	int64
1	age	1000 non-null	int64
2	policy_number	1000 non-null	int64
3	policy_bind_date	1000 non-null	datetime64[ns]
4	policy_state	1000 non-null	object
5	policy_csl	1000 non-null	object
6	policy_deductable	1000 non-null	int64
7	policy_annual_premium	1000 non-null	float64
8	umbrella_limit	1000 non-null	int64
9	insured_zip	1000 non-null	int64
10	insured_sex	1000 non-null	object
11	insured_education_level	1000 non-null	object
12	insured_occupation	1000 non-null	object
13	insured_hobbies	1000 non-null	object
14	insured_relationship	1000 non-null	object
15	capital-gains	1000 non-null	int64
16	capital-loss	1000 non-null	int64
17	incident_date	1000 non-null	datetime64[ns]

18	incident_type	1000 non-null	object
19	collision_type	1000 non-null	object
20	incident_severity	1000 non-null	object
21	authorities_contacted	1000 non-null	object
22	incident_state	1000 non-null	object
23	incident_city	1000 non-null	object
24	incident_location	1000 non-null	object
25	incident_hour_of_the_day	1000 non-null	int64
26	number_of_vehicles_involved	1000 non-null	int64
27	property_damage	1000 non-null	object
28	bodily_injuries	1000 non-null	int64
29	witnesses	1000 non-null	int64
30	police_report_available	1000 non-null	object
31	total_claim_amount	1000 non-null	int64
32	injury_claim	1000 non-null	int64
33	property_claim	1000 non-null	int64
34	vehicle_claim	1000 non-null	int64
35	auto_make	1000 non-null	object
36	auto_model	1000 non-null	object
37	auto_year	1000 non-null	int64
38	fraud_reported	1000 non-null	object
39	_c39	0 non-null	float64

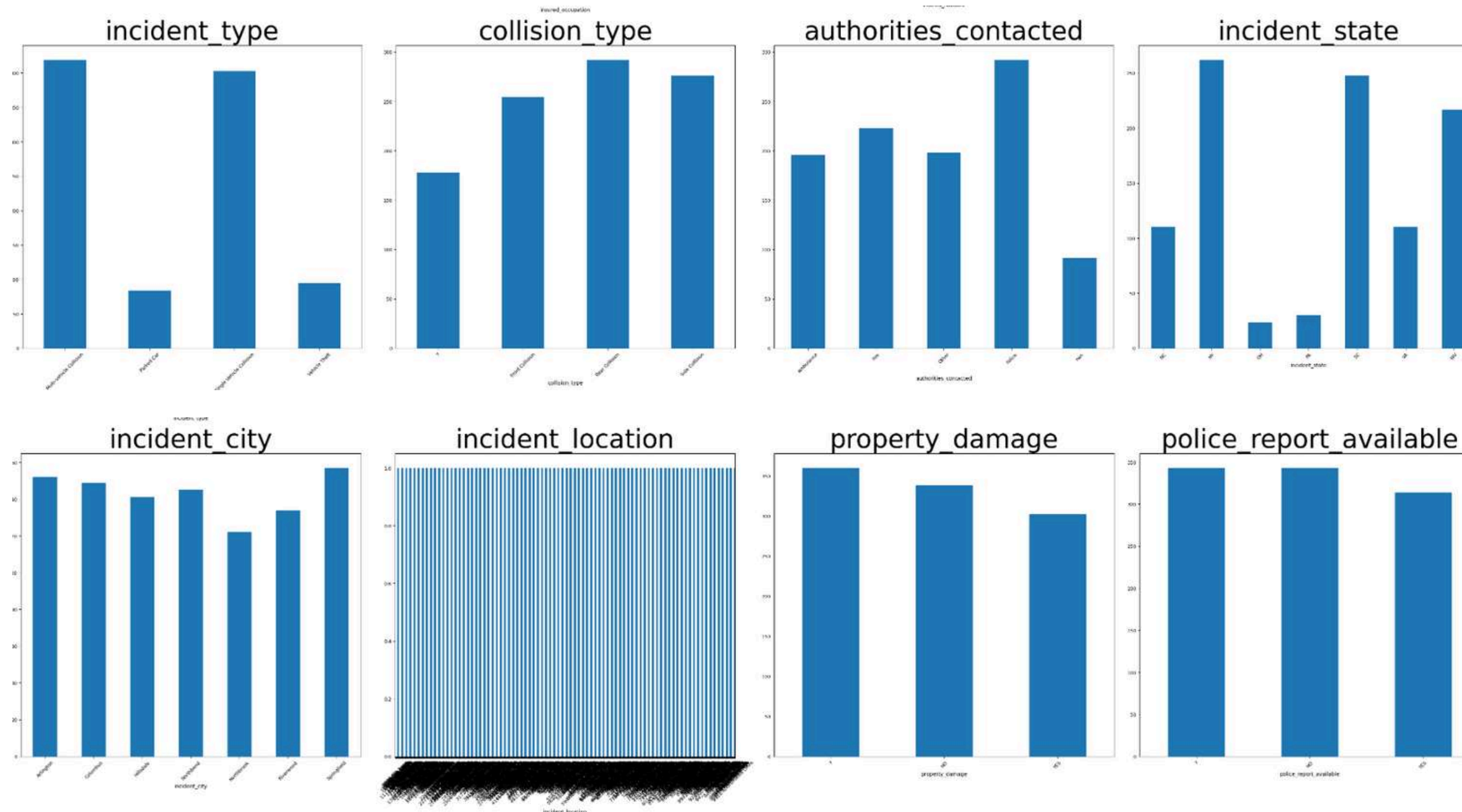
dtypes: datetime64[ns](2), float64(2), int64(17), object(19)





# Eksplorasi Data

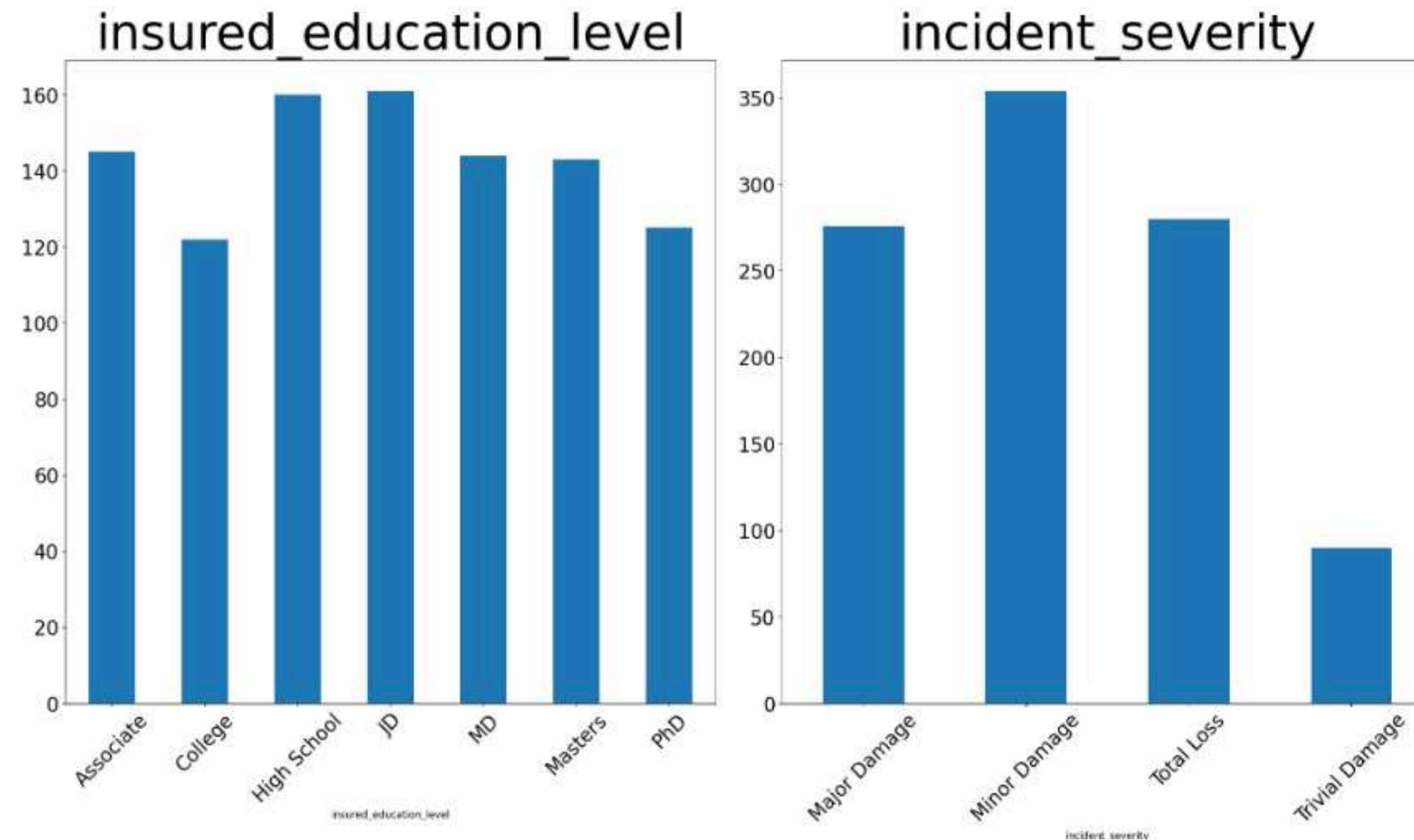
## Plot Fitur Nominal



insured_relationship	count
husband	170
not-in-family	175
other-relative	178
non-child	185
unmarried	140
wife	155

# Eksplorasi Data

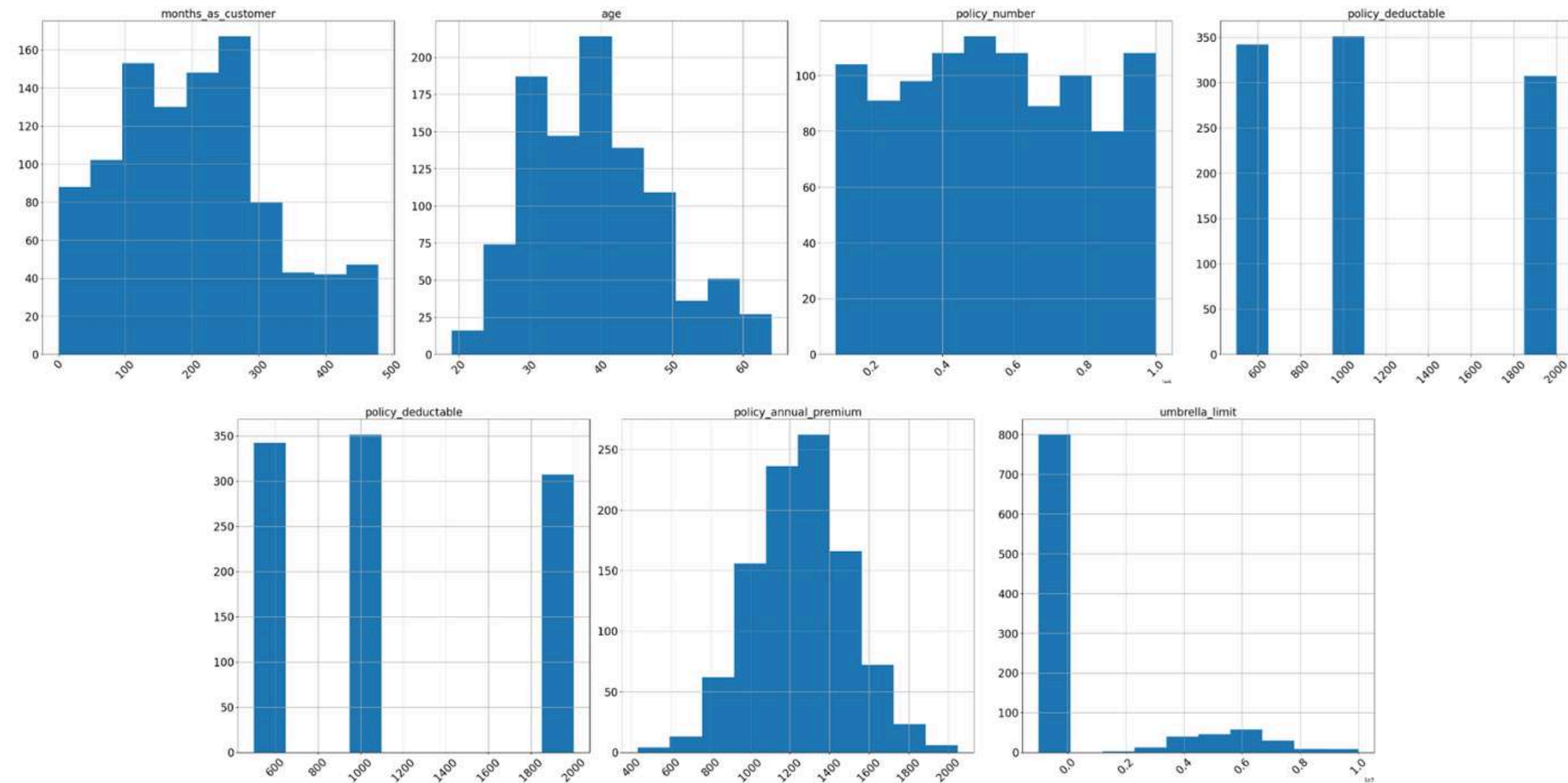
## Plot Fitur Ordinal





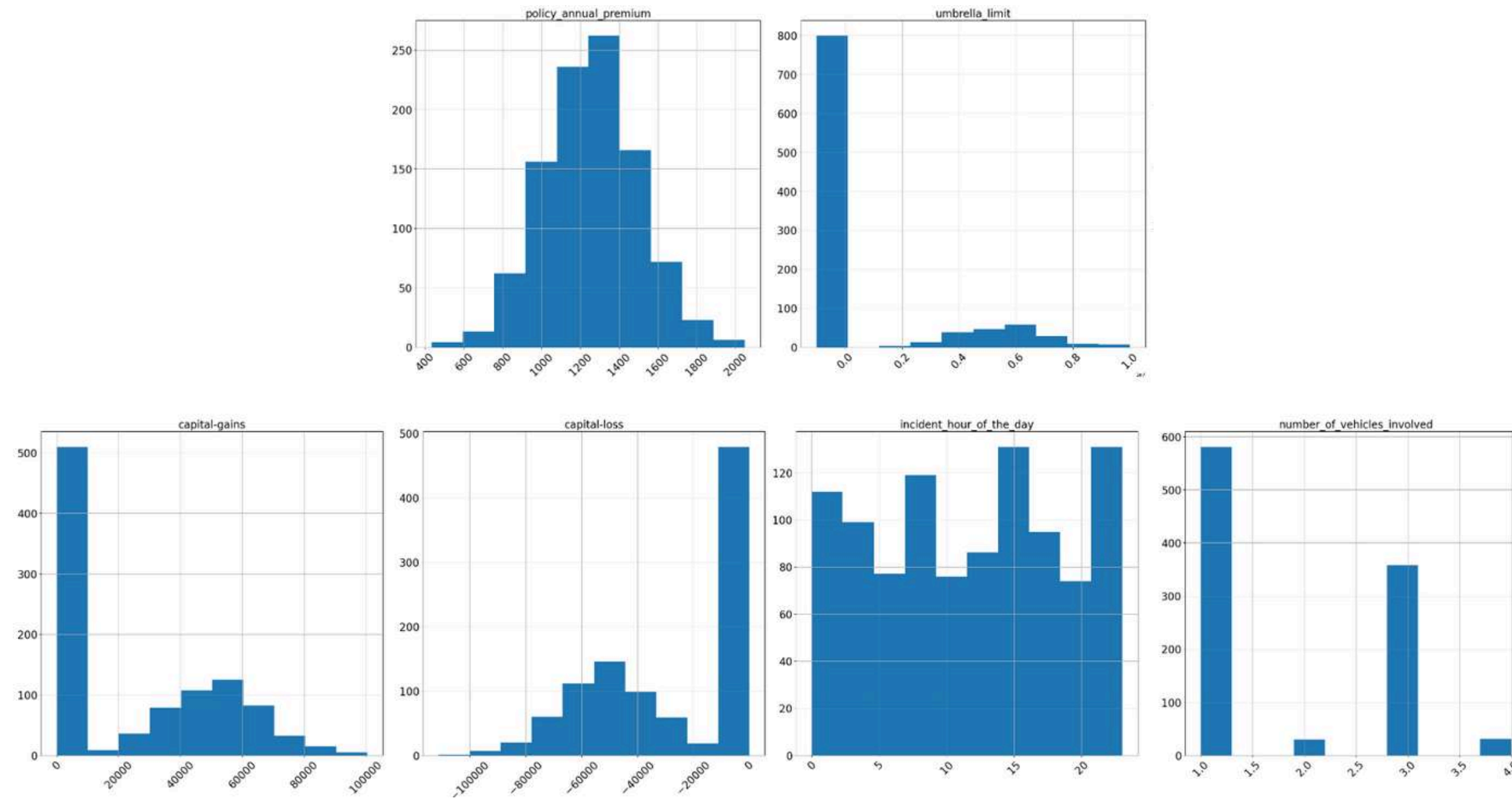
# Eksplorasi Data

## Plot Fitur Numerik



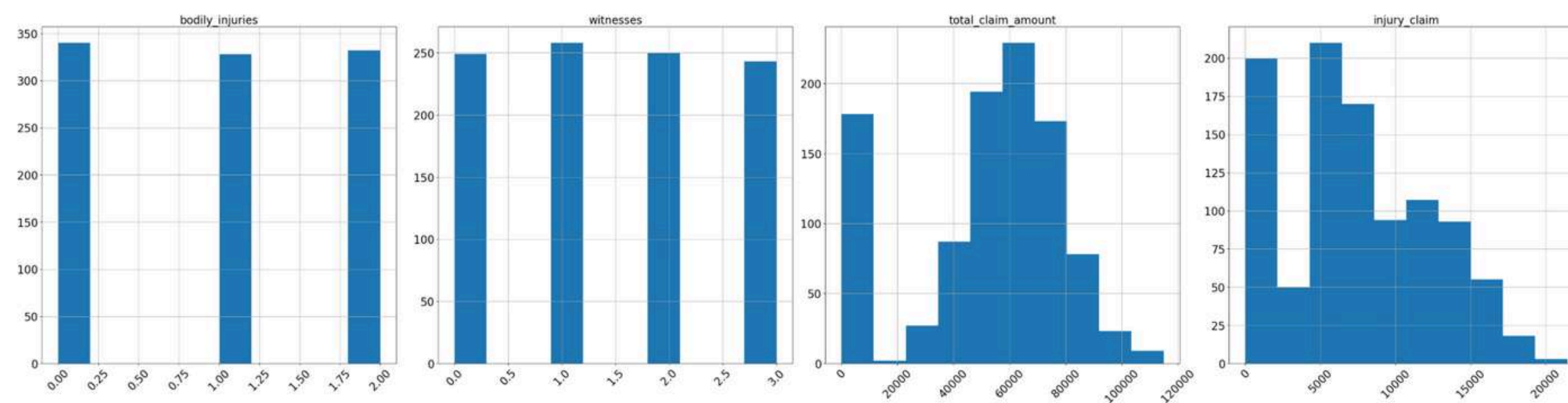
# Eksplorasi Data

## Plot Fitur Numerik



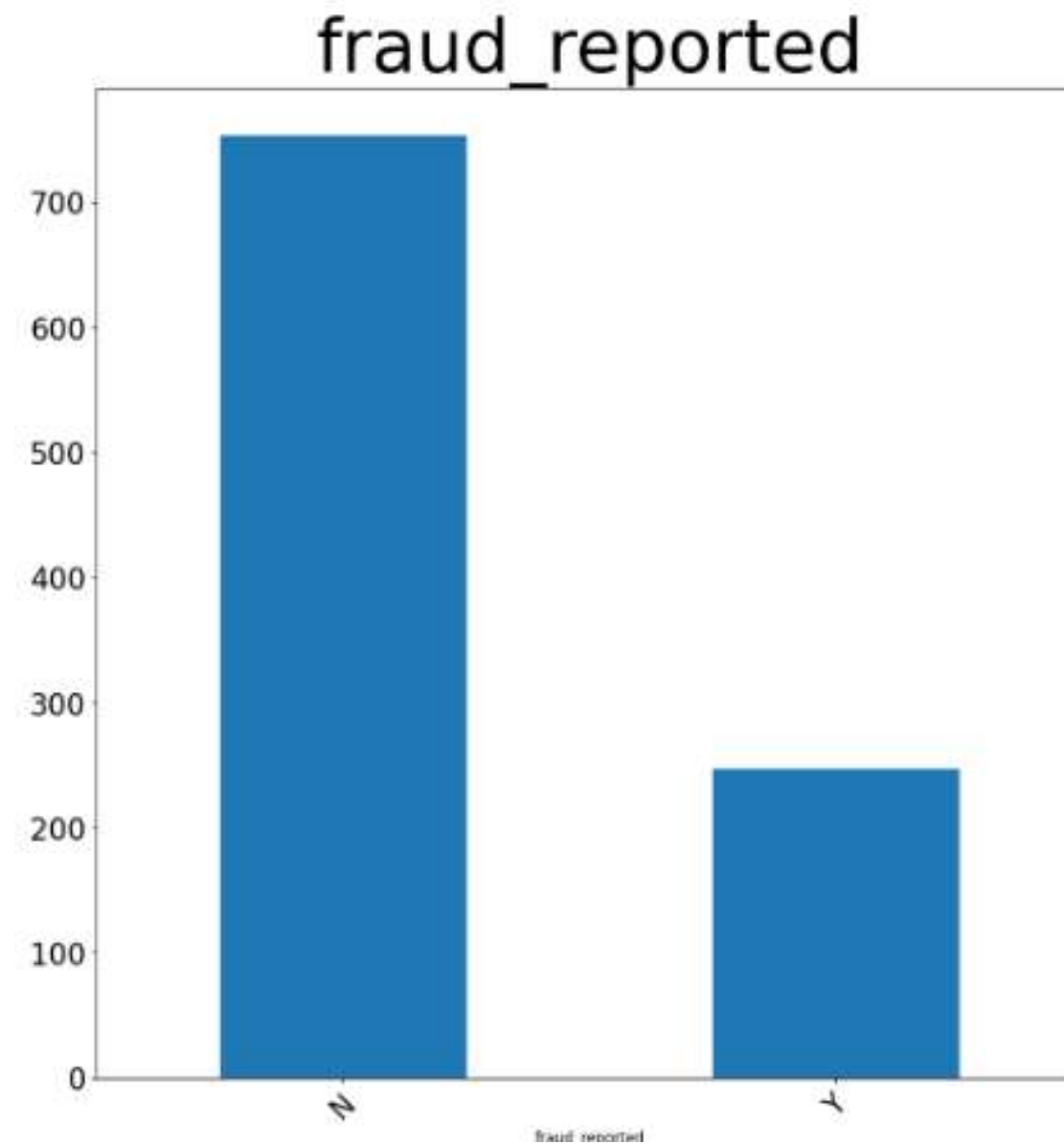
# Eksplorasi Data

## Plot Fitur Numerik



# Eksplorasi Data

## Plot Fitur Nominal (Respon)





# Eksplorasi Data

## Missing Values

```
Columns containing NaN: ['authorities_contacted', '_c39']
```

```
Counts of NaN in each column:
```

```
authorities_contacted      91
```

```
_c39                      1000
```

```
dtype: int64
```

```
Entries with NaN in each column:
```

```
authorities_contacted: [4, 13, 27, 37, 51, 52, 57, 69, 78, 81, 83, 88, 92, 95,
```

```
_c39: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
```

# Eksplorasi Data

## Missing Values

```
Columns containing '?': ['collision_type', 'property_damage', 'police_report_available']
```

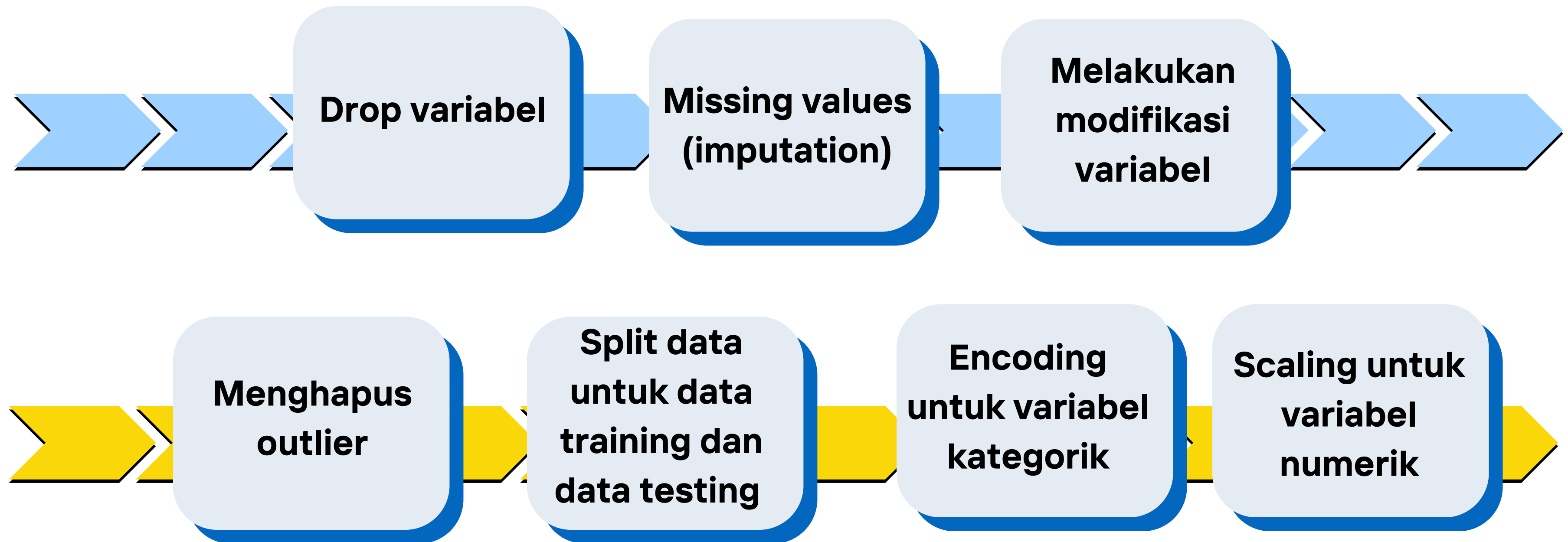
```
Counts of '?' in each column:
```

```
collision_type          178  
property_damage         360  
police_report_available 343  
dtype: int64
```

```
Entries with '?' in each column:
```

```
collision_type: [1, 4, 13, 26, 27, 37, 48, 51, 52, 54, 57, 69, 78, 81, 82, 83, 88, 92, 95, 98, 99, 103  
property_damage: [1, 3, 6, 7, 16, 19, 21, 23, 36, 38, 39, 41, 48, 50, 51, 52, 56, 70, 79, 81, 84, 87,  
police_report_available: [1, 6, 9, 10, 20, 21, 23, 27, 30, 31, 33, 38, 40, 41, 44, 45, 46, 49, 51, 53,
```

# Pre-Processing Data



# Drop Variabel

Variabel yang Dibuang	Alasan
policy_number	Hanya berupa <i>identifier</i> , tidak memberikan informasi apapun
insured_zip	Hanya berupa <i>identifier</i> dan terdapat 995 nilai unik dari 1000 entri.
insured_relationship	Informasi kurang jelas karena berisi status pernikahan di suatu entri tetapi status kekerabatan di entri lainnya.
incident_location	Setiap entri berisi nilai yang berbeda (unik). Sebagai gantinya, variabel incident_city tetap digunakan.
auto_make	variabel auto_model merupakan subset dari auto_make dan memberikan informasi lebih detail.

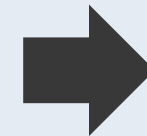


# Drop Variabel

Menghapus kolom 'policy\_number', 'insured\_zip', 'insured\_relationship', 'incident\_location', 'auto\_make'

policy\_number

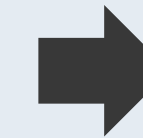
age	policy_number	policy_bind_date
48	521585	2014-10-17
42	342868	2006-06-27
29	687698	2000-09-06
41	227811	1990-05-25
44	367455	2014-06-06
39	104594	2006-10-12



age	policy_bind_date
48	2014-10-17
42	2006-06-27
29	2000-09-06
41	1990-05-25
44	2014-06-06

insured\_zip

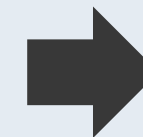
umbrella_limit	insured_zip	insured_sex
0	466132	MALE
5000000	468176	MALE
5000000	430632	FEMALE
6000000	608117	FEMALE
6000000	610706	MALE
0	478456	FEMALE



umbrella_limit	insured_sex
0	MALE
5000000	MALE
5000000	FEMALE
6000000	FEMALE
6000000	FEMALE
6000000	MALE

auto\_make

vehicle_claim	auto_make	auto_model
52080	Saab	92x
3510	Mercedes	E400
23100	Dodge	RAM
50720	Chevrolet	Tahoe
4550	Accura	RSX



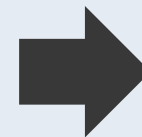
vehicle_claim	auto_model
52080	92x
3510	E400
23100	RAM
50720	Tahoe
4550	RSX

# Drop Variabel

Menghapus kolom 'policy\_number', 'insured\_zip', 'insured\_relationship', 'incident\_location', 'auto\_make'

insured\_relationship

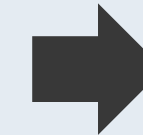
insured_hobbies	insured_relationship	capital-gains
sleeping	husband	53300
reading	other-relative	0
board-games	own-child	35100
board-games	unmarried	48900
board-games	unmarried	66000



incident_date	incident_type
2015-01-25	Single Vehicle Collision
2015-01-21	Vehicle Theft
2015-02-22	Multi-vehicle Collision
2015-01-10	Single Vehicle Collision
2015-02-17	Vehicle Theft

incident\_location

incident_city	incident_location	incident_hour_of_the_day
Columbus	9935 4th Drive	5
Riverwood	6608 MLK Hwy	8
Columbus	7121 Francis Lane	7
Arlington	6956 Maple Drive	5
Arlington	3041 3rd Ave	20



incident_city	incident_hour_of_the_day
Columbus	5
Riverwood	8
Columbus	7
Arlington	5
Arlington	20

# Mengolah Missing Values

## BEFORE

	authorities_contacted	_c39
419	NaN	NaN
420	NaN	NaN
421	NaN	NaN
422	NaN	NaN
423	NaN	NaN

## PROCESS

## AFTER

	authorities_contacted
419	No Contact
420	No Contact
421	No Contact
422	No Contact
423	No Contact

- Data NaN pada authorities\_contacted aslinya bernilai "None" tetapi terbaca sebagai NaN karena bahasa pemrograman python. Data NaN diubah menjadi "No Contact" agar dataset tetap bisa diproses.
- Label \_c39 bukan merupakan label untuk sebuah kolom, melainkan informasi tambahan pada dataset bahwa terdapat 39 kolom (variabel) pada dataset. Oleh karena itu, kolom \_c39 harus dihapus.

# Mengolah Missing Values

## BEFORE

	collision_type	property_damage	police_report_available
2	Front Collision	?	?
3	Side Collision	NO	?
4	Front Collision	NO	?
419	?	NO	NO
423	?	?	?

## PROCESS

## AFTER

	collision_type	property_damage	police_report_available
2	Front Collision	NO	NO
3	Side Collision	NO	NO
4	Front Collision	NO	NO
419	Rear Collision	NO	NO
423	Rear Collision	NO	NO

Dilakukan imputasi variabel kategorik dengan mengganti *missing values* dengan modus dari masing-masing variabel.



# Menghapus Outlier

```
Original DataFrame shape: (1000, 34)
DataFrame shape after outlier removal: (988, 34)
Removed 12 rows.
Indices of removed rows: [79, 168, 359, 412, 430, 449, 532, 552, 602, 647, 651, 850]
```

- Terdapat 12 baris outlier yang terdapat pada dataframe.
- Menghapus 12 baris outlier pada data
- Data frame berubah dari memiliki 1000 baris dan 34 kolom data menjadi 988 baris dan 34 kolom data

# Memodifikasi Variabel

	policy_bind_date	incident_date	Duration
0	1997-11-20	2015-01-22	6272
1	1994-05-27	2015-01-11	7534
2	1991-02-08	2015-01-19	8746
3	2000-02-18	2015-01-18	5448
4	1992-04-04	2015-01-09	8315

Duration merupakan jumlah hari antara kapan polis terbit dan insiden terjadi.

Setelah diperoleh variabel Duration, variabel policy\_bind\_date dan incident\_date dihapus.

# Splitting Dataset

```
X_train.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
239	40	500	1463.95	0	
143	34	500	1442.27	0	
342	49	500	1722.95	0	
147	34	1000	1275.81	0	
184	38	1000	1437.53	0	

```
y_train.head()
```

	fraud_reported_Y
935	False
598	False
237	False
264	False
94	True

```
X_test.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
81	25	500	920.30	5000000	
54	35	500	1261.28	0	
253	41	2000	1312.75	0	
210	35	500	1433.24	0	
259	44	2000	1655.79	0	

```
y_test.head()
```

	fraud_reported_Y
544	True
357	False
565	True
315	True
654	False

# Encoding

## BEFORE

policy_state	
0	OH
1	IL
2	IN
3	IL
4	IL

authorities_contacted	
4	Ambulance
243	Fire
494	No Contact
675	Other
999	Police

insured_sex	
3	MALE
4	MALE
11	FEMALE
12	FEMALE

P  
R  
O  
C  
E  
S  
S

## AFTER

	policy_state_IN	policy_state_OH
0	False	True
1	False	False
2	True	False
3	False	False
4	False	False

	authorities_contacted_Fire	authorities_contacted_No Contact	authorities_contacted_Other	authorities_contacted_Police
4	False	False	False	False
243	True	False	False	False
494	False	True	False	False
675	False	False	True	False
999	False	False	False	True

insured_sex_MALE	
3	True
4	True
11	False
12	False

## ONE-HOT ENCODER

One-Hot Encoder merupakan mekanisme pengolahan variabel kategorik nominal di mana sebuah n-level variabel dikonversi menjadi (n-1) variabel dummy dengan nilai True dan False.

Baseline diperoleh ketika semua variabel dummy bernilai False.





# Encoding

## BEFORE

	insured_education_level	incident_severity
0	MD	Total Loss
1	High School	Major Damage
5	JD	Major Damage
6	Associate	Minor Damage
7	PhD	Total Loss
14	Masters	Major Damage
40	College	Minor Damage
434	College	Trivial Damage

P  
R  
O  
C  
E  
S  
S

## AFTER

	insured_education_level	incident_severity
0	6	4
1	3	3
5	5	3
6	1	2
7	7	4
14	4	3
40	2	2
434	2	1

## ORDINAL ENCODER

Ordinal Encoder merupakan mekanisme pengolahan variabel kategorik ordinal di mana nilai dari variabel tersebut diubah menjadi angka sesuai dengan urutan.



# Scaling

BEFORE

```
X_train.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
239	40	500	1463.95	0	
143	34	500	1442.27	0	
342	49	500	1722.95	0	
147	34	1000	1275.81	0	
184	38	1000	1437.53	0	

```
X_test.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
81	25	500	920.30	5000000	
54	35	500	1261.28	0	
253	41	2000	1312.75	0	
210	35	500	1433.24	0	
259	44	2000	1655.79	0	

P  
R  
O  
C  
E  
S  
S

AFTER

```
X_train.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
0.498956	0.466667	0.000000	0.646738	0.111111	
0.298539	0.333333	0.000000	0.631593	0.111111	
0.713987	0.666667	0.000000	0.827672	0.111111	
0.306889	0.333333	0.333333	0.515306	0.111111	
0.384134	0.422222	0.333333	0.628282	0.111111	

```
X_test.head()
```

months_as_customer	age	policy_deductable	policy_annual_premium	umbrella_limit	
0.169102	0.133333	0.0	0.266951	0.666667	
0.112735	0.355556	0.0	0.505156	0.111111	
0.528184	0.488889	1.0	0.541112	0.111111	
0.438413	0.355556	0.0	0.625285	0.111111	
0.540710	0.555556	1.0	0.780755	0.111111	



# SMOTE

## Mengolah Data Tidak Seimbang

BEFORE

count	
fraud_reported_Y	
False	596
True	194

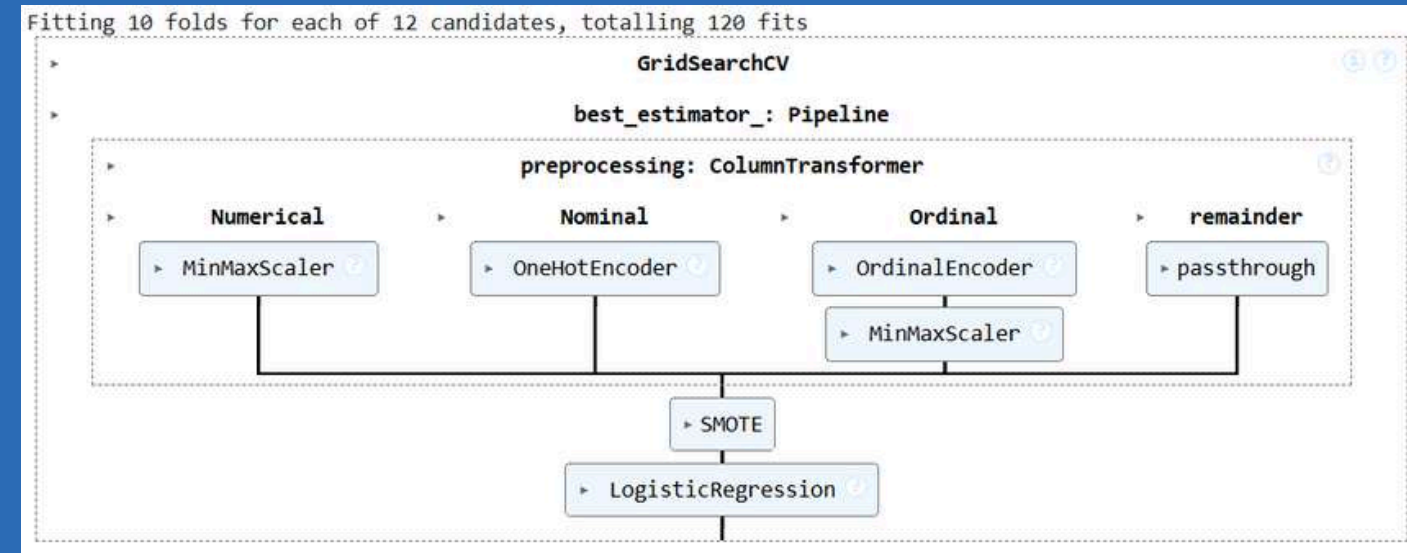
PROCESS

AFTER

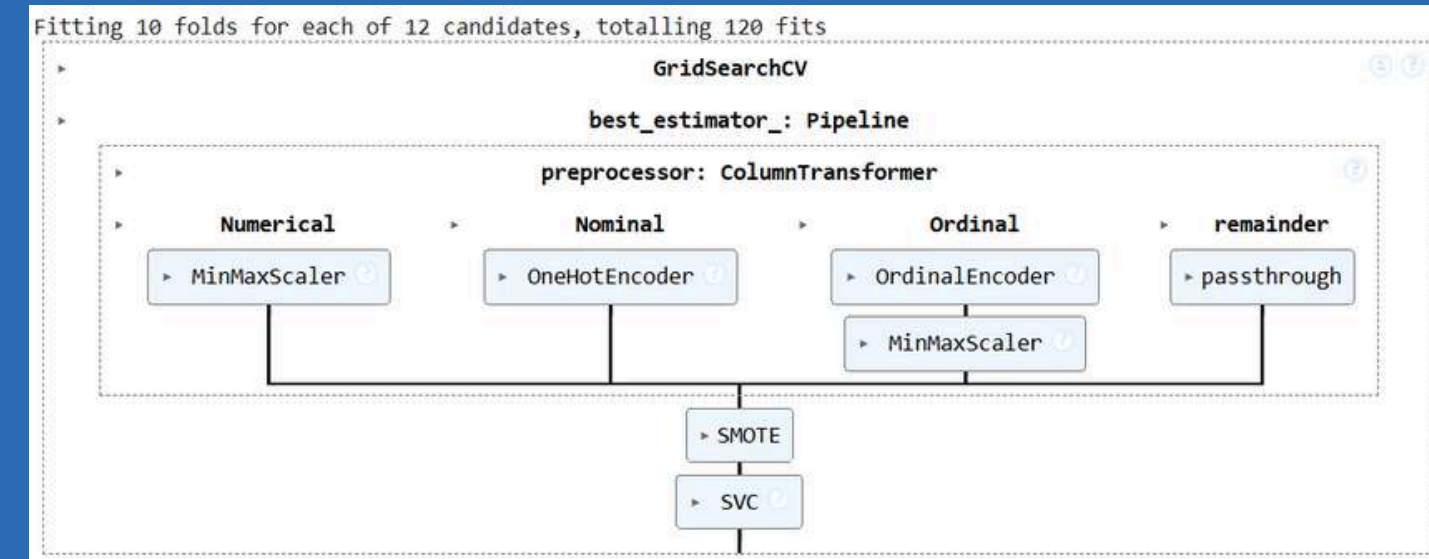
count	
fraud_reported_Y	
False	596
True	596

# Modelling: Pipeline

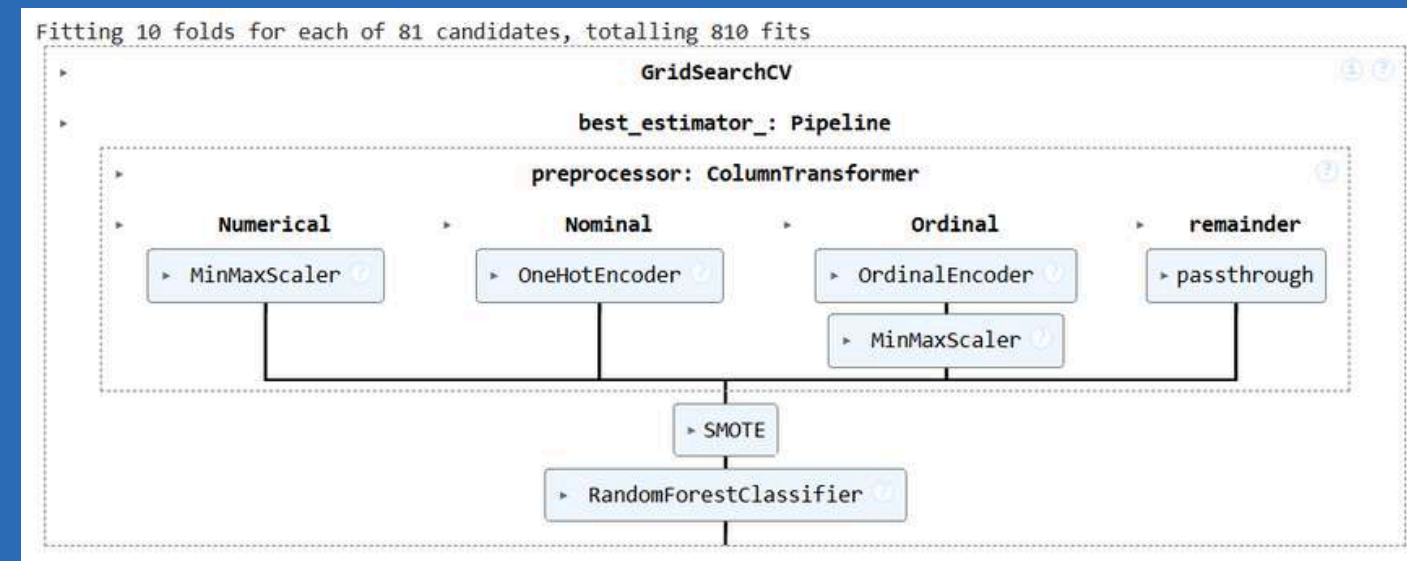
## Logistic Regression



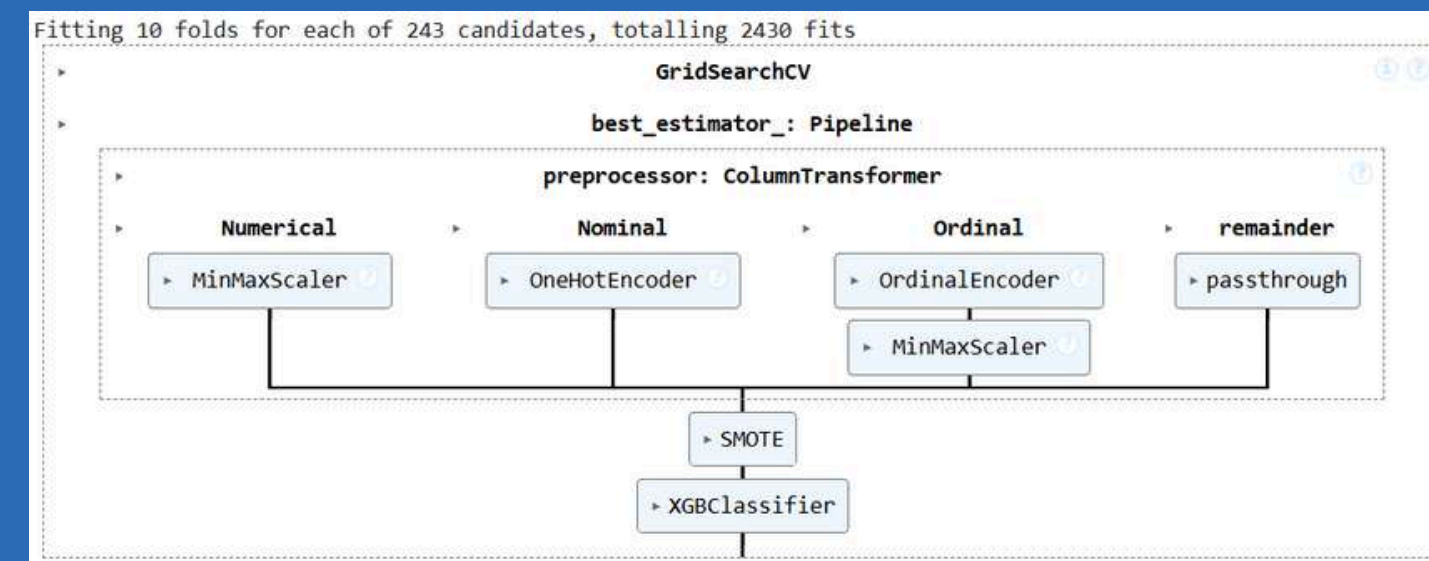
## SVC



## Random Forest



## XGBoost





# Modelling: Optimasi Parameter

## Logistic Regression

Parameter	Nilai Optimal
C	1
penalty	l1
solver	saga

## Support Vector Classification (SVC)

Parameter	Nilai Optimal
C	0.1
class_weight	balanced
degree	2
gamma	1
kernel	rbf
max_iter	500

# Modelling: Optimasi Parameter

**Random Forest**

Parameter	Nilai Optimal
max_depth	6
min_samples_leaf	50
min_samples_split	150

**XGBoost**

Parameter	Nilai Optimal
eta	0.05
gamma	0
lambda	9
max depth	9

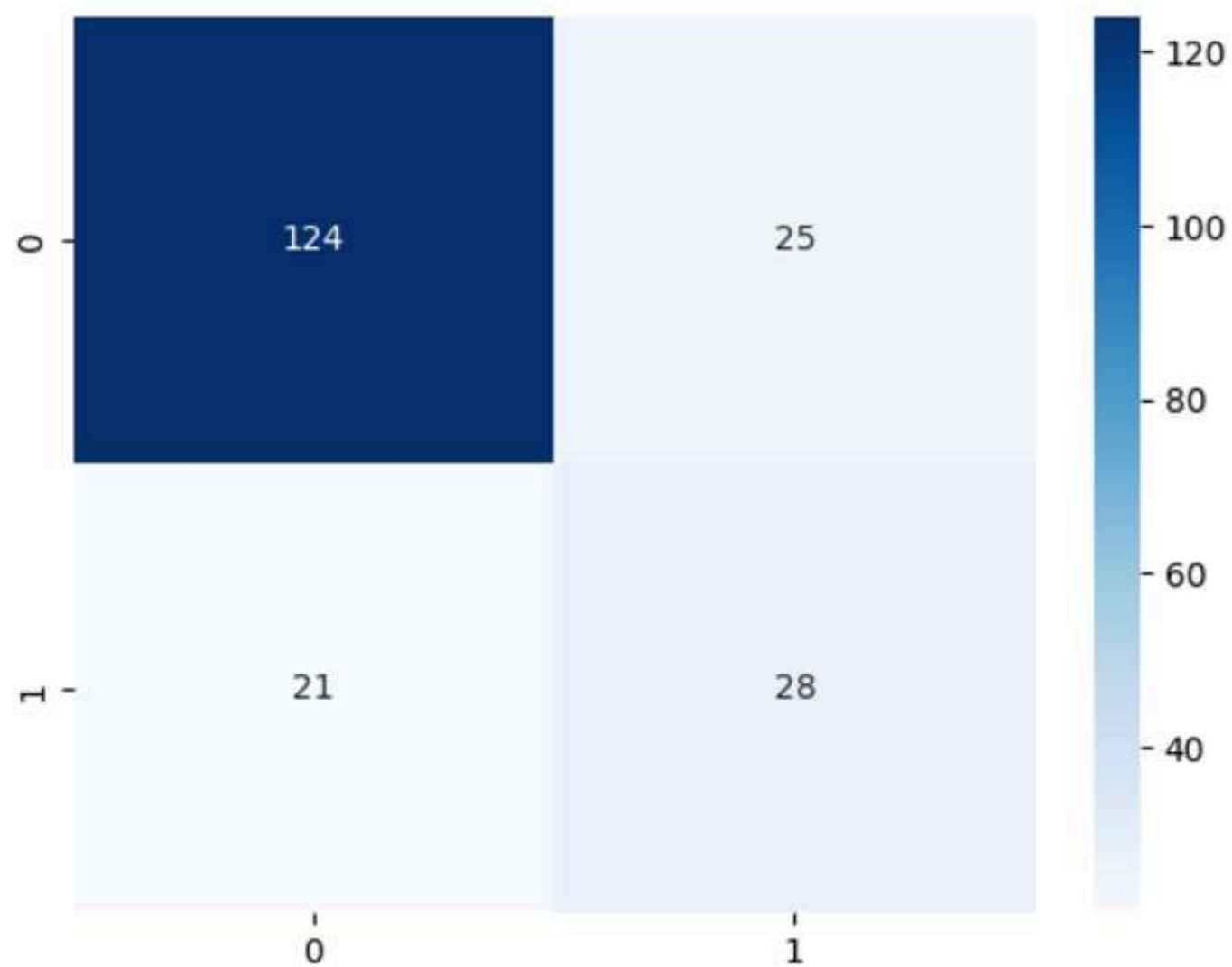
# Modelling: Kinerja Model

Model	Random Forest	Logistic Regression	Support Vector Classification (SVC)	XGBoost
AUC-ROC	0.64306259416	0.72427983539	0.5	0.7898917956444322
Specificity	0.6577181208	0. 0.673611111111112	1	0.8322147651006712
Sensitiity	0.551020408	0.59259259259259	0	0.5714285714285714

Berdasarkan perbandingan dari nilai AUC-ROC, Specificity, dan Sensitivity keempat model, didapat model terbaik adalah **XGBoost**.

# Best Model: Confussion Matrix

Confusion Matrix of XGBoost Model



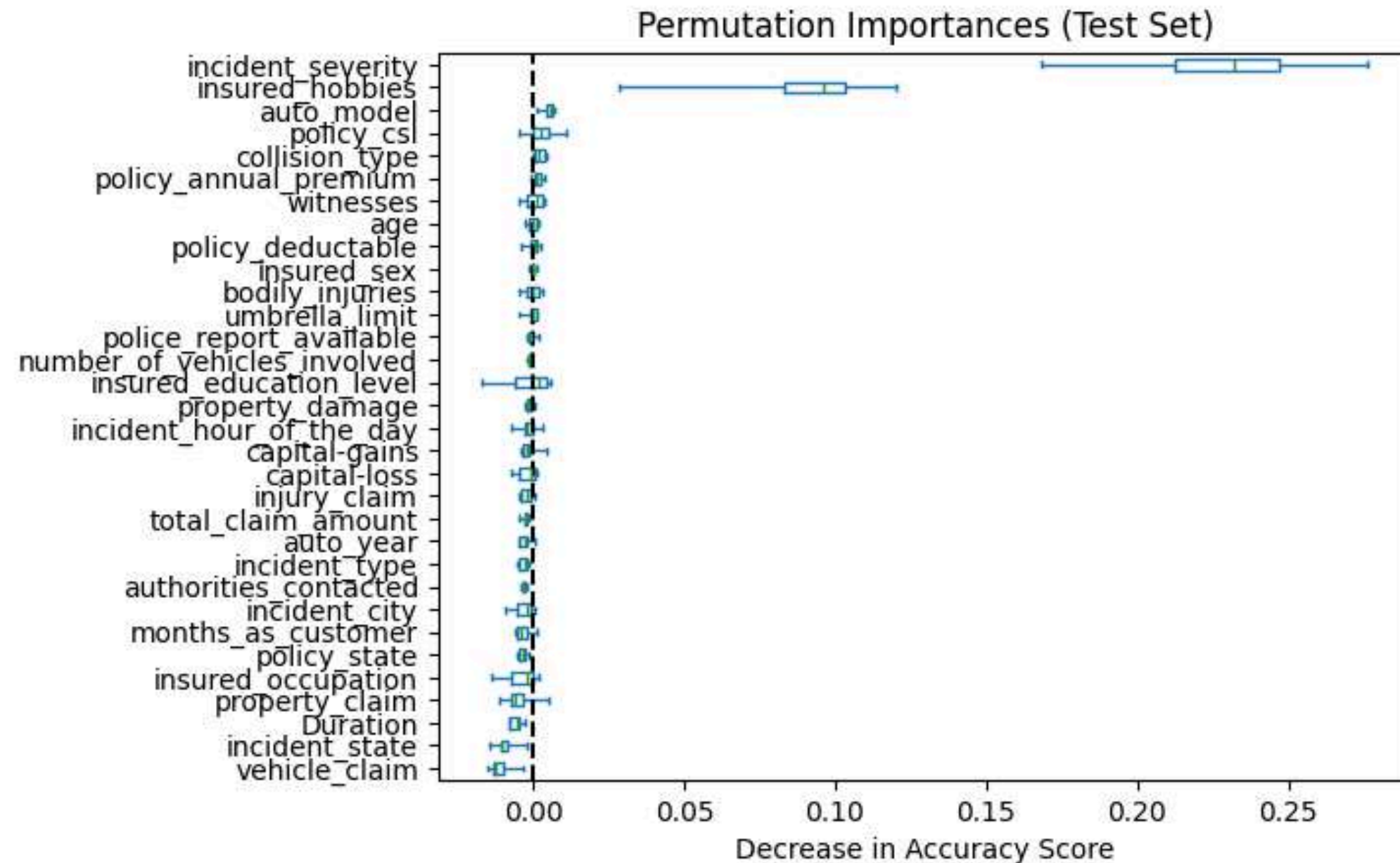
**Specificity**

$$\frac{TN}{TN + FP} = \frac{28}{28 + 21} = 0,5714$$

**Sensitivity**

$$\frac{TP}{TP + FN} = \frac{124}{124 + 25} = 0.8322$$

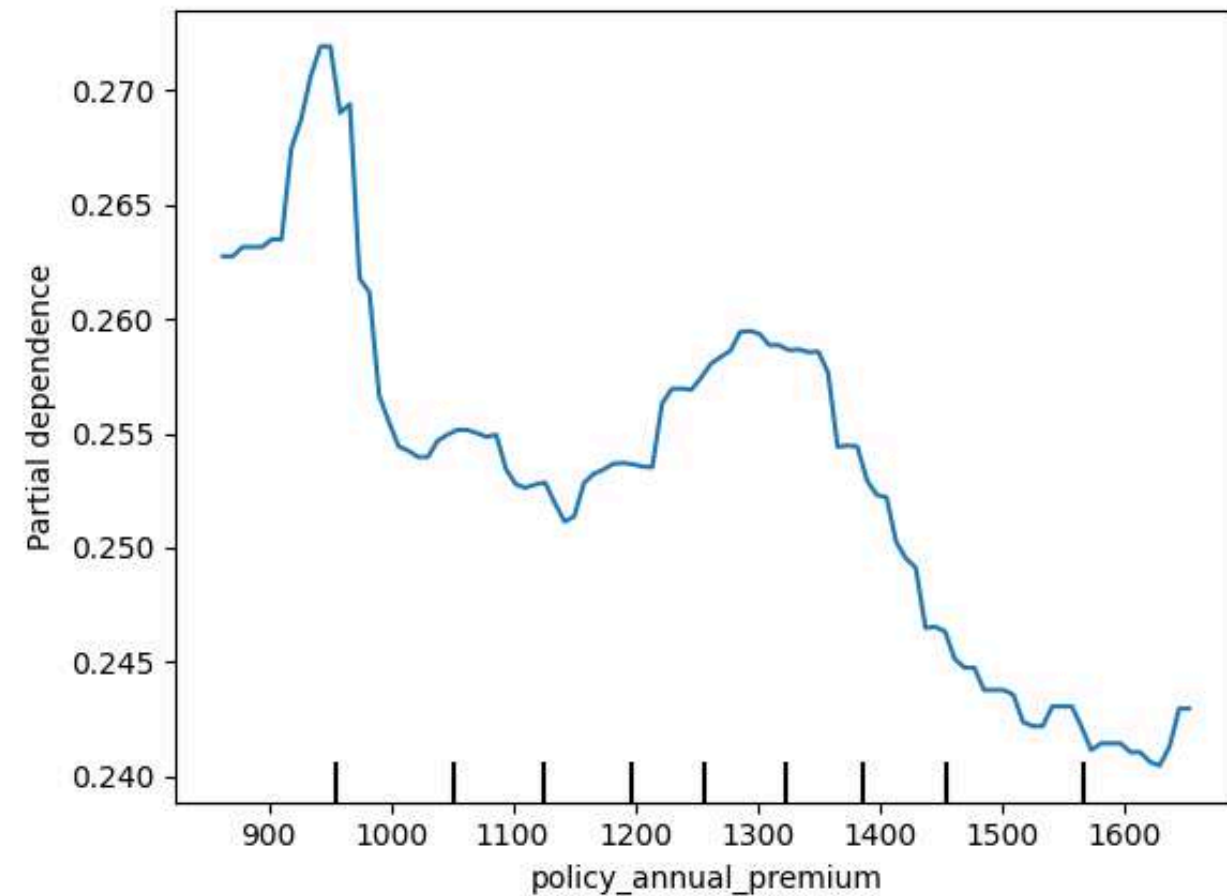
# Best Model: Feature Importance



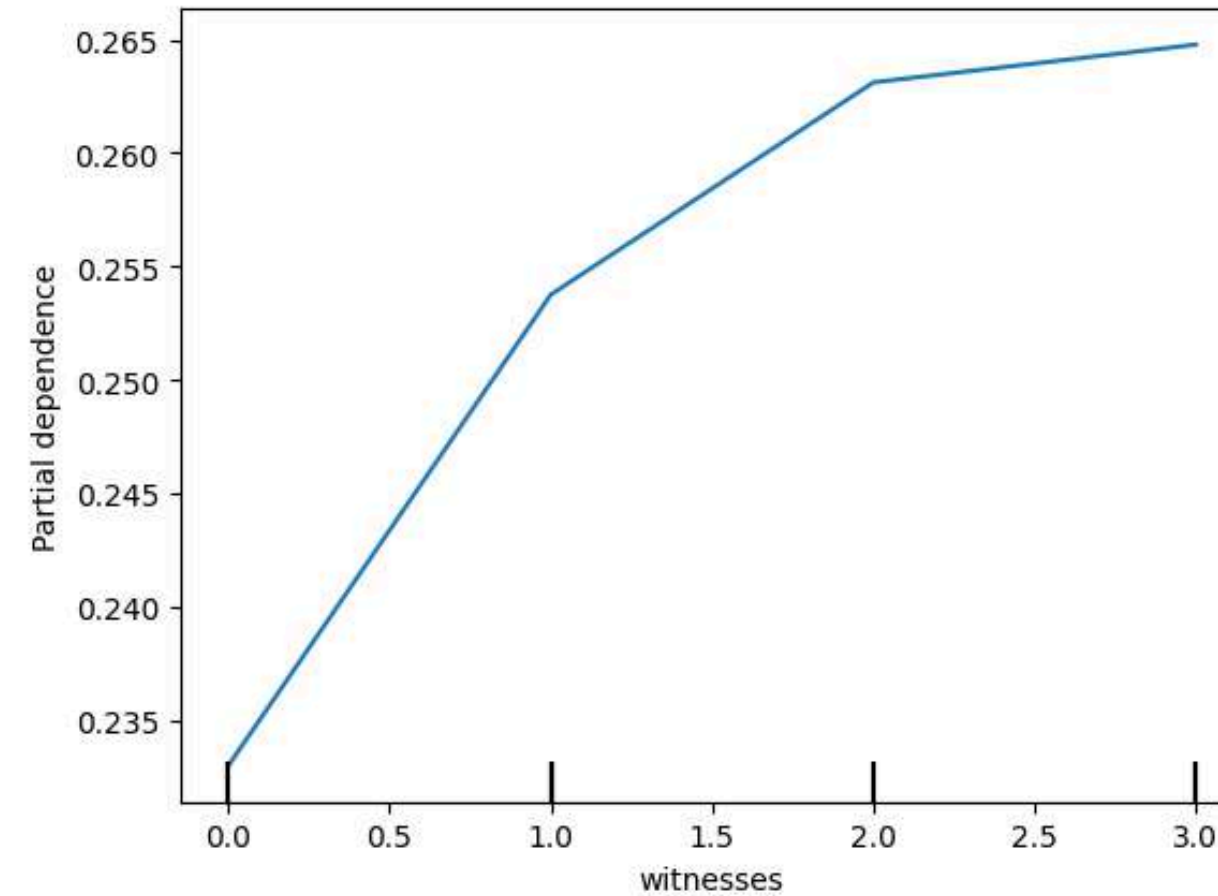
Dilihat dari feature importance, fitur "incident\_severity", "insured\_hobbies", dan "auto\_model" akan meningkatkan akurasi prediksi model.



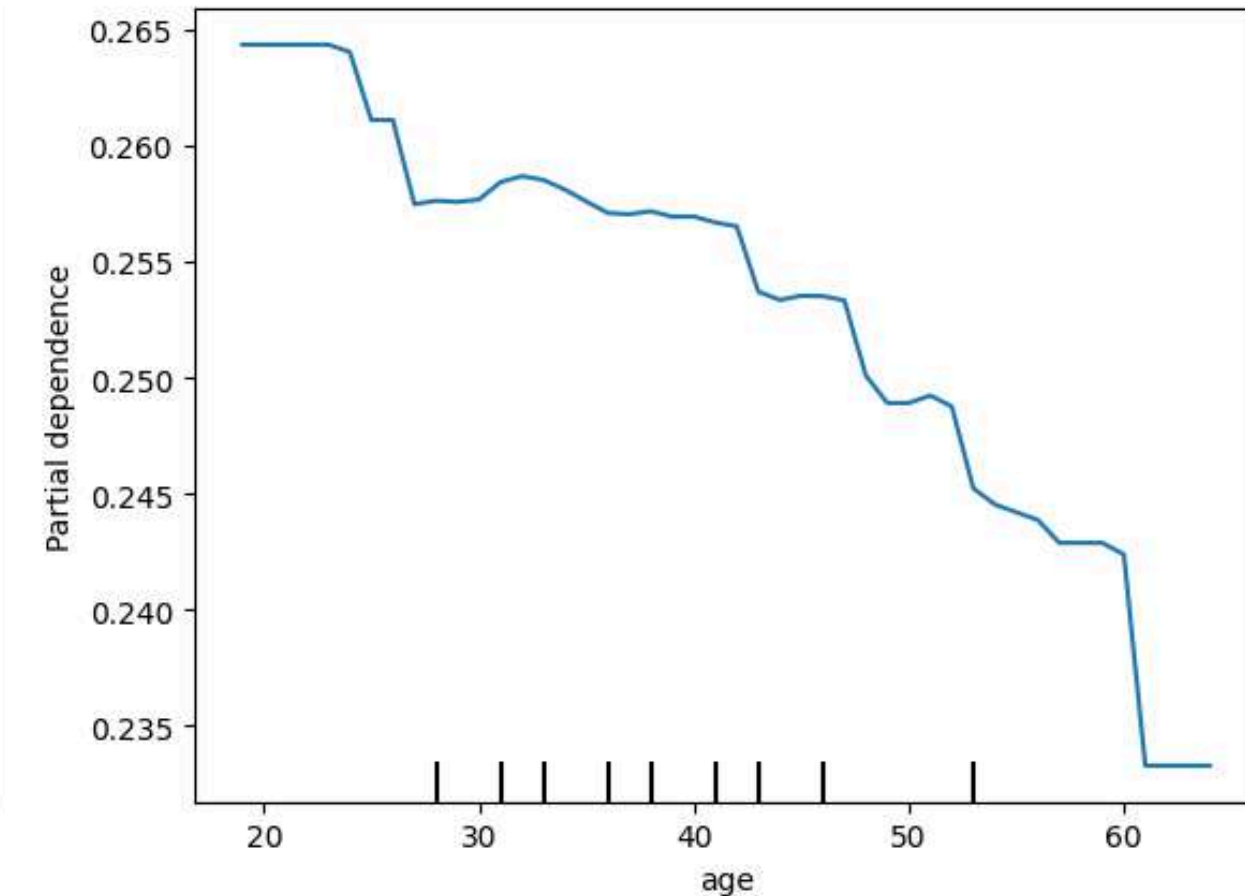
# Best Model: PDP



Berdasarkan grafik, peluang terjadi fraud akan meningkat pada saat nilai premi dalam rentang 900–970 dan 1140–1290 dan akan cenderung menurun pada saat interval nilai premi lainnya

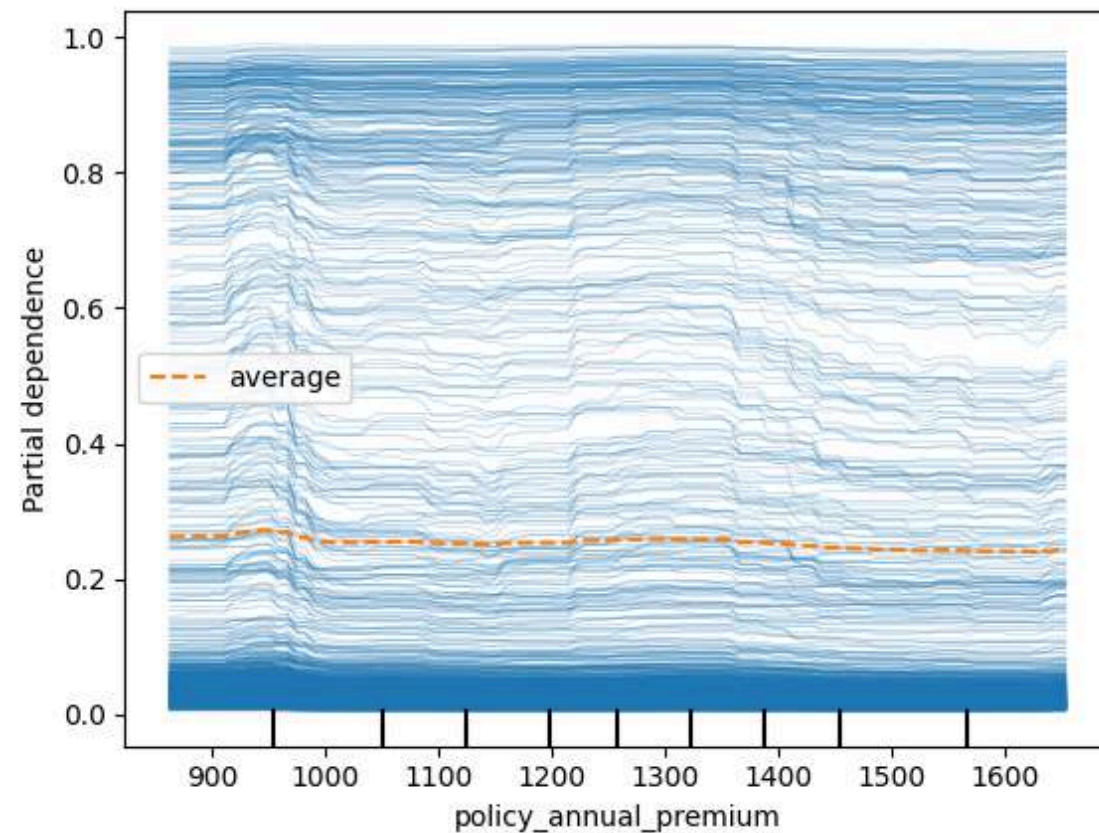


Berdasarkan dari grafik, semakin banyak jumlah saksi yang melaporkan maka semakin tinggi probabilitas terjadinya fraud. (witnesses memiliki hubungan positif dengan probabilitas fraud)

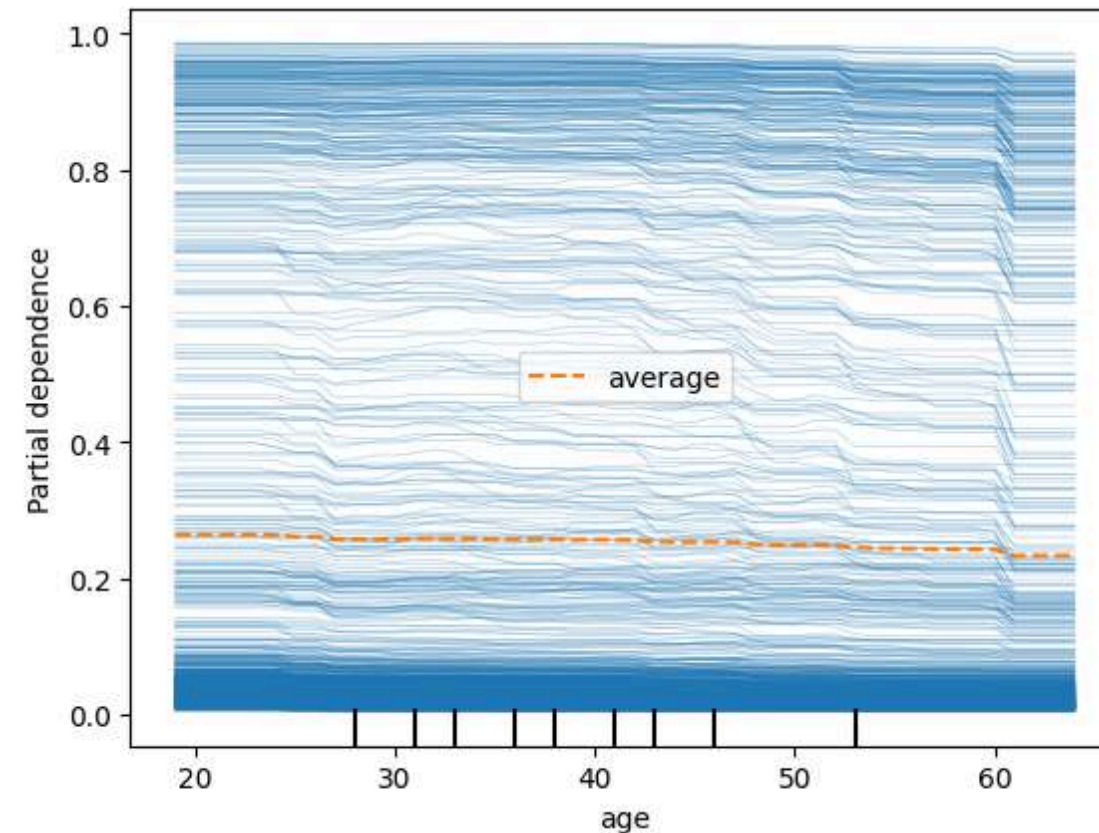


Berdasarkan dari grafik, semakin tua (semakin besar usia) pemegang polis, maka probabilitas terjadinya fraud cenderung menurun. (age memiliki hubungan negatif dengan probabilitas fraud)

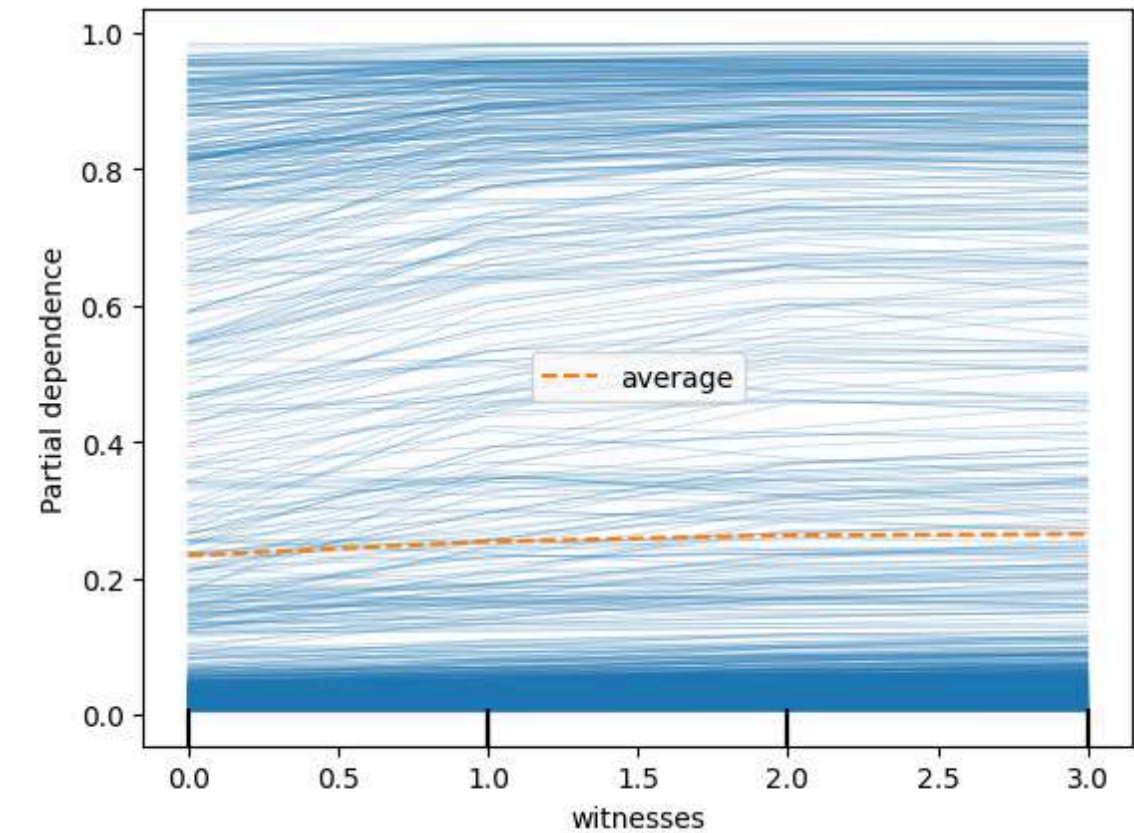
# Best Model : ICE



Dilihat dari grafik, secara keseluruhan fitur ini memiliki pengaruh yang rendah dengan kecenderungan di area non fraud, meskipun ada beberapa kasus outlier. Peluang terjadi fraud akan meningkat pada beberapa interval, namun akan menurun dan kembali normal pada interval setelahnya



Dilihat dari grafik, secara keseluruhan fitur ini memiliki pengaruh yang sedikit rendah dengan kecenderungan di area non fraud, meskipun ada beberapa kasus outlier. Secara keseluruhan semakin tinggi usia, peluang terjadi fraud akan semakin menurun



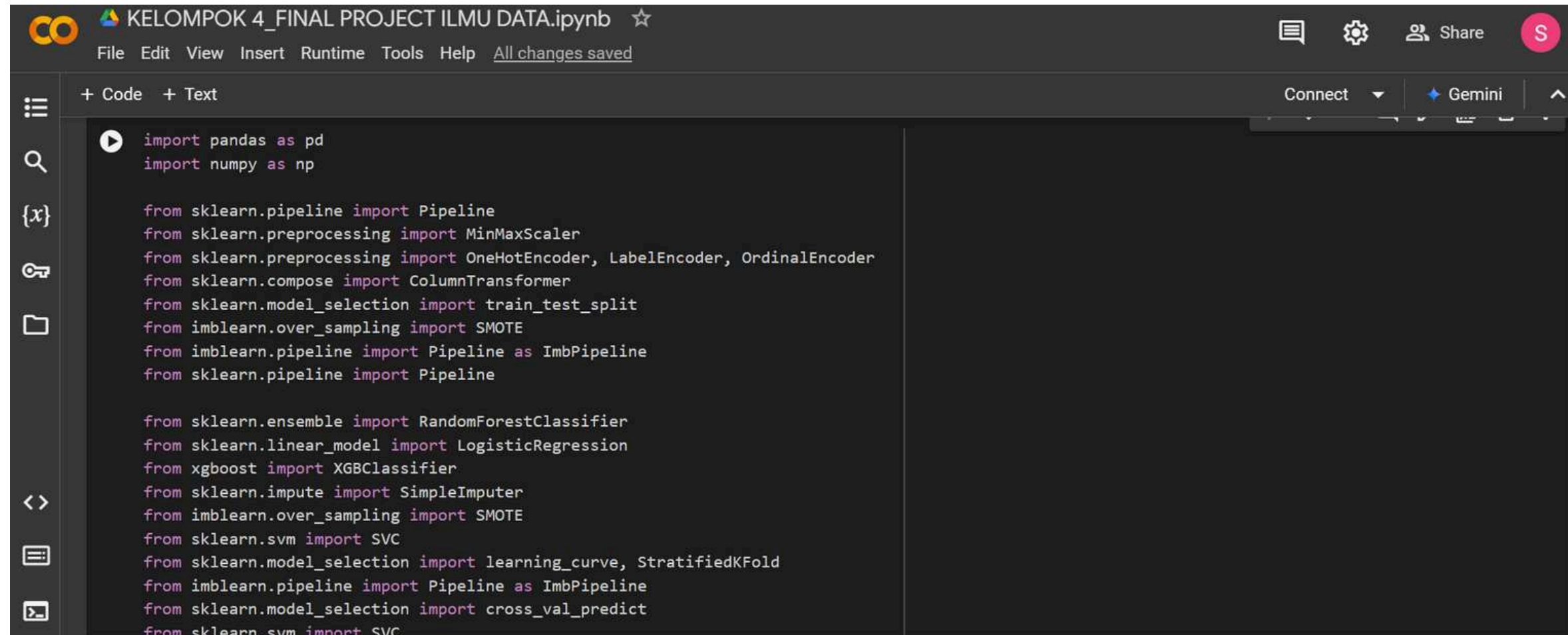
Dilihat dari grafik, secara keseluruhan fitur ini memiliki pengaruh yang cukup baik dengan kecenderungan di area non fraud, meskipun ada beberapa kasus outlier. Secara keseluruhan semakin banyak jumlah saksi yang melapor, peluang terjadinya fraud akan semakin tinggi

Melalui hasil evaluasi, didapatkan bahwa di antara pemodelan Random Forest, Logistic Regression, Support Vector Clasification, dan XGBoost, didapatkan bahwa pemodelan yang terbaik adalah pemodelan berdasarkan XGBoost dengan AUC-ROC Score sebesar 0.7898917956444322.

Interpretasi dari model XGBoost diberikan dalam 3 aspek, yaitu Feature Importance, PDP, dan ICE. Berdasarkan aspek Feature Importance, didapatkan bahwa variabel "incident\_severity" adalah variabel yang paling memberi pengaruh dalam meningkatkan akurasi dalam prediksi model.



# Lampiran



```
import pandas as pd
import numpy as np

from sklearn.pipeline import Pipeline
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import OneHotEncoder, LabelEncoder, OrdinalEncoder
from sklearn.compose import ColumnTransformer
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
from imblearn.pipeline import Pipeline as ImbPipeline
from sklearn.pipeline import Pipeline

from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from xgboost import XGBClassifier
from sklearn.impute import SimpleImputer
from imblearn.over_sampling import SMOTE
from sklearn.svm import SVC
from sklearn.model_selection import learning_curve, StratifiedKFold
from imblearn.pipeline import Pipeline as ImbPipeline
from sklearn.model_selection import cross_val_predict
from sklearn.svm import SVC
```

**Code untuk pemrosesan data dapat diakses melalui tautan berikut:**

**<https://colab.research.google.com/drive/1lZO-f-aZN8udykXmsgLb26nqoFZdLPPNb?usp=sharing#scrollTo=VGKUBgL28BT1>**

# Daftar Pustaka

Hendri Murfi. (2024). Ilmu Data. Lecture Notes.

Muhammad Adli Rahmat Solihin. (2024). Analisis Kinerja Model Approximating XGBoost untuk Deteksi Fraud Klaim Asuransi.





**TERIMA KASIH**