



# **PEMODELAN PENGARUH BENCANA BANJIR DAN KONDISI EKONOMI REGIONAL TERHADAP PDRB INDONESIA MENGGUNAKAN MACHINE LEARNING**

- 1. Annisa Sekartierra Mulyanto**  
**(22/494177/PA/21257)**
- 2. Mahardi Nalendra Syafa**  
**(22/502515/PA/21558)**



**DISUSUN OLEH ADAMANTINE**  
*AWESOME DATA MINING  
MAHARDI AND TIERRA PROJECT*



# Latar Belakang

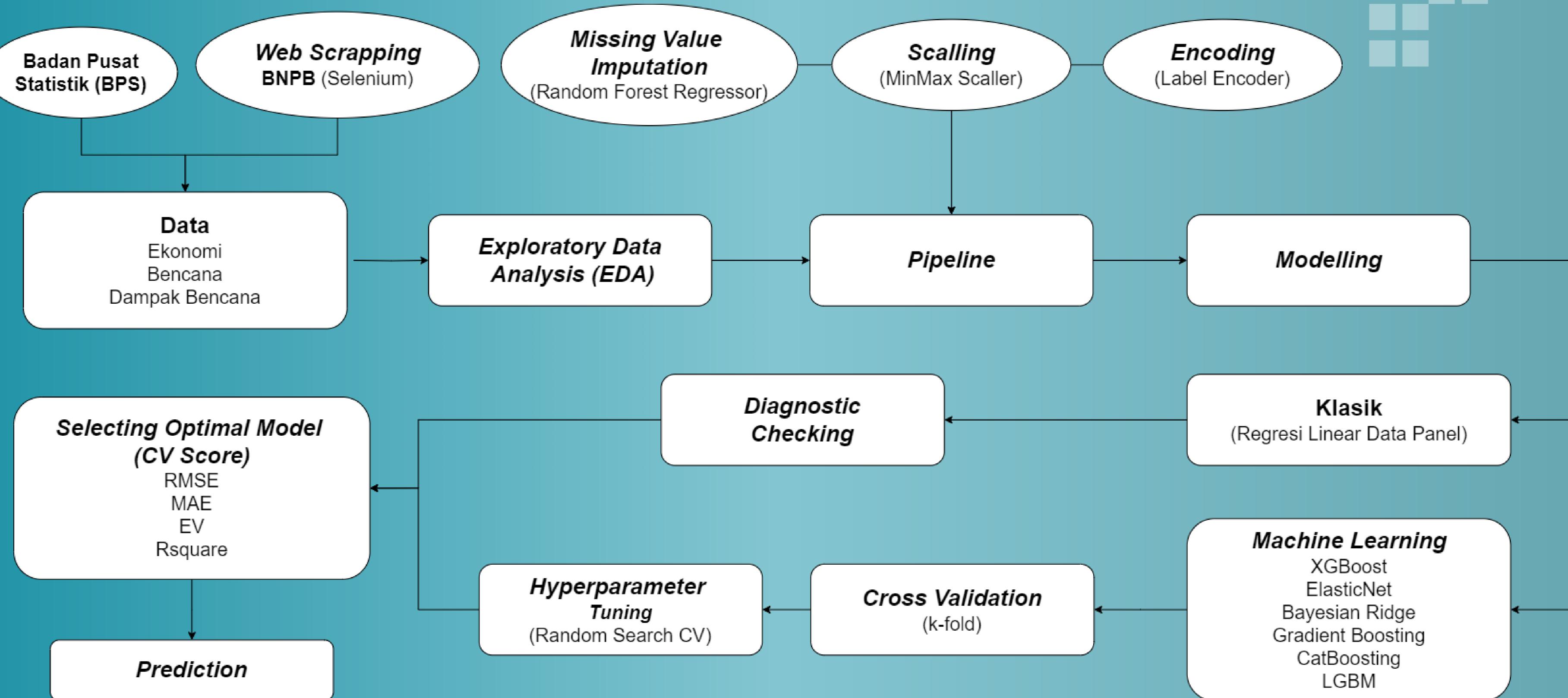
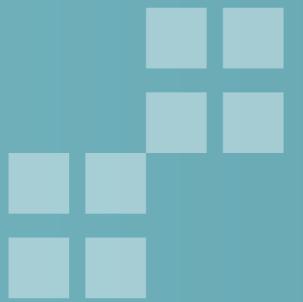
Tidak bisa dipungkiri bahwa Indonesia merupakan daerah rawan bencana. Faktanya, bedasarkan World Risk Report, Indonesia menduduki **peringkat kedua negara yang paling berisiko bencana di dunia**. Salah satu bencana alam yang sering terjadi di Indonesia adalah banjir. Pada tahun 2022, tercatat **banjir** adalah **bencana yang paling sering terjadi** yakni sebanyak 1520 kejadian. Bedasarkan data Badan Nasional Penanggulangan Bencana, jumlah jiwa terpapar risiko bencana banjir tersebar dibeberapa pulau di Indonesia dengan jumlah melebihi 170 juta jiwa dan nilai aset terpapar **melebihi Rp 750 triliun**.

Angka kerugian ini diduga berdampak pada indeks ekonomi suatu daerah. Dengan kata lain, kejadian bencana banjir diduga berdampak pada indeks ekonomi daerah. Oleh karena itu, penelitian ini akan melihat **pengaruh kejadian banjir** terhadap indeks ekonomi (**PDRB**) tiap provinsi dan membangun **model prediktif** untuk mengestimasikan PDRB berdasarkan kejadian banjir di Indonesia.

## Tujuan

- ➡ Membentuk **model** yang baik dalam memprediksi PDRB tiap provinsi di Indonesia berdasarkan variabel bencana, variabel akibat bencana, dan variabel ekonomi.
- ➡ Mengidentifikasi variabel-variabel apa saja yang **berpengaruh kuat** terhadap PDRB suatu provinsi.

# Metodologi



# Data



Data yang digunakan berada pada rentang waktu **2015-2023** pada 34 provinsi di Indonesia. Integrasi data dilakukan dengan mengumpulkan data dari **BPS** dan **BNPB** dimana beberapa prosesnya menggunakan **web scrapping**.

Terdapat **15** variabel numerik dan **1** kategorik yang dibagi menjadi tiga bagian:

## Variabel Akibat Bencana

- Korban meninggal
- Korban hilang
- Korban luka-luka
- Korban menderita
- Korban mengungsi
- Bangunan terdampak
- Indeks bencana banjir

## Variabel Bencana

- Curah hujan
- Frekuensi curah hujan sedang
- Frekuensi curah hujan lebat
- Jumlah bencana



## Variabel Ekonomi

- Provinsi
- Jumlah penduduk
- Tingkat pengangguran terbuka
- Realisasi investasi penanaman modal luar negeri
- Realisasi investasi penanaman modal dalam negeri

# Pre-Processing

## *Train test splitting*

Train data

**2015-2022**

Test data

**2023**

## Di Pipeline....

### 1. Imputasi Missing Value

Dilakukan dengan menggunakan **ItrativeImputer()** khususnya algoritma **RandomForestRegressor()** dimana proses imputasi dipisahkan bedasarkan kelompok variabel.

### 2. Data transformation

Dilakukan dengan menggunakan **MinMaxScaler()** untuk seluruh variabel numerik pada data.

$$v'_i = \frac{v_i - \min}{\max - \min} (\text{new\_max} - \text{new\_min}) + \text{new\_min}$$

$v'_i$  adalah data setelah transformasi dan  $v_i$  adalah data sebelum transformasi

### 3. Encoding

Dilakukan encoding untuk variabel provinsi menggunakan **label encoding**.

# Metode Klasik:

## Analisis Regresi Linear Data Panel

Gabungan antara data runtun waktu dan data cross section disebut data panel.  
(Rosadi, 2022)

Persamaan umum regresi data panel:

$$Y_{it} = X'_{t,i}\beta + C_i + d_t + u_{t,i}$$

$Y_{i,t}$  : PDRB untuk provinsi ke- $i$  dan tahun ke- $t$

$X_{t,i}$  : Variabel independen untuk provinsi ke- $i$  dan tahun ke- $t$

$\beta_k$  : Koefisien variabel independen ke- $k$ ,  $k = 1, 2, \dots$

$C_i$  : Konstanta yang bergantung pada individu ke- $i$

$d_t$  : Konstanta yang bergantung pada waktu ke- $t$

$u_{t,i}$  : Komponen galat dari komponen runtun waktu dan kali-silang.

### Model

- **Fixed Effect**

One Way

$$Y_{it} = X'_{t,i}\beta + C_i + u_{t,i}$$

$$Y_{it} = X'_{t,i}\beta + d_t + u_{t,i}$$

Two Way

$$Y_{it} = X'_{t,i}\beta + C_i + d_t + u_{t,i}$$

- **Random Effect**

$$Y_{i,t} = X'_{i,t}\beta + v_{i,t}$$

di mana  $v_{i,t} = C_i + d_t + \varepsilon_{i,t}$ .

Memperlakukan efek spesifik entitas yang tidak teramat sebagai **efek acak** dan **tidak berkorelasi dengan variabel penjelas**.

### Pengujian

#### **Uji Hausman**

Fixed Effect vs Random Effect

#### **Uji Breusch Pagan**

Efek individu/waktu/two-way

#### **Uji Wald**

Overall & Partial

### Diagnostic Checking

- **Normalitas Residual**

Kolmogorov-Smirnov Test

- **No Autokorelasi / Korelasi Serial**

Uji Breusch-Godfrey/Wooldridge

- **Homoskedastisitas Residual**

Uji Breusch-Pagan



# Metode Machine Learning: *Gradient Boosting Regressor*

Gradient Boosting Regressor adalah teknik yang membangun model secara bertahap dan menggabungkan mereka (ensemble) untuk meningkatkan performa model untuk masalah regresi.

## Step 1

Inisialisasi model dengan single leaf

## Step 2

Membuat trees sebanyak M dalam loop

1. Menghitung residual dengan menderivatifkan loss function (*gradient*)

$$L = \frac{1}{2} (actual - predicted)^2$$

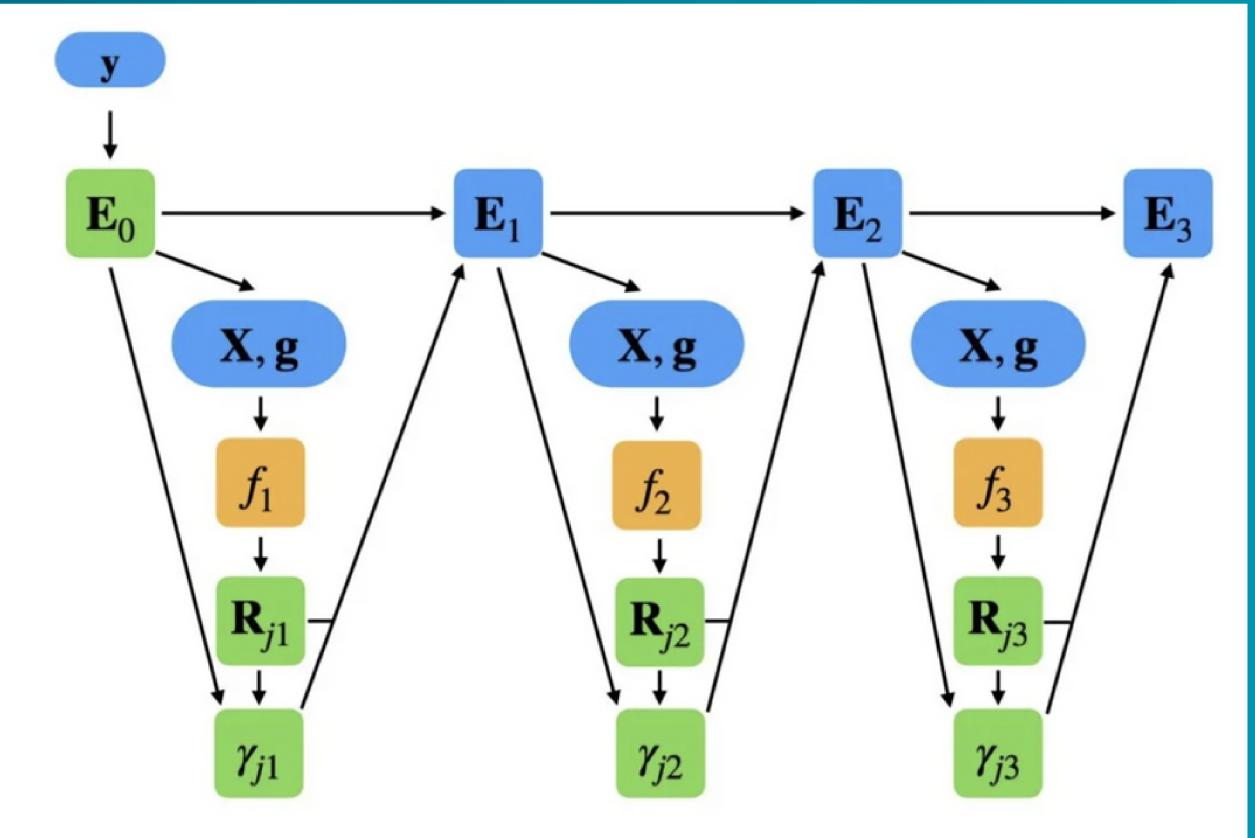
2. Membangun *base learner* berupa decision tree (*f*) dengan variabel target adalah residual (*R*)

3. Menghitung nilai gamma yang memminimumkan

$$\gamma_{jm} = \operatorname{argmin}_\gamma \sum_{x_i \in R_{ij}} L(y_i, E_{m-1}(x_i) + \gamma)$$

4. Mengupdate prediksi dengan mempertimbangkan learning rate (*v*)

$$E_m(x) = E_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$



## Metrik evaluasi

### RMSE

$$\sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

### MAE

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

### EV

$$1 - \frac{Var(y - \hat{y})}{Var(y)}$$

### R square

$$1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$$

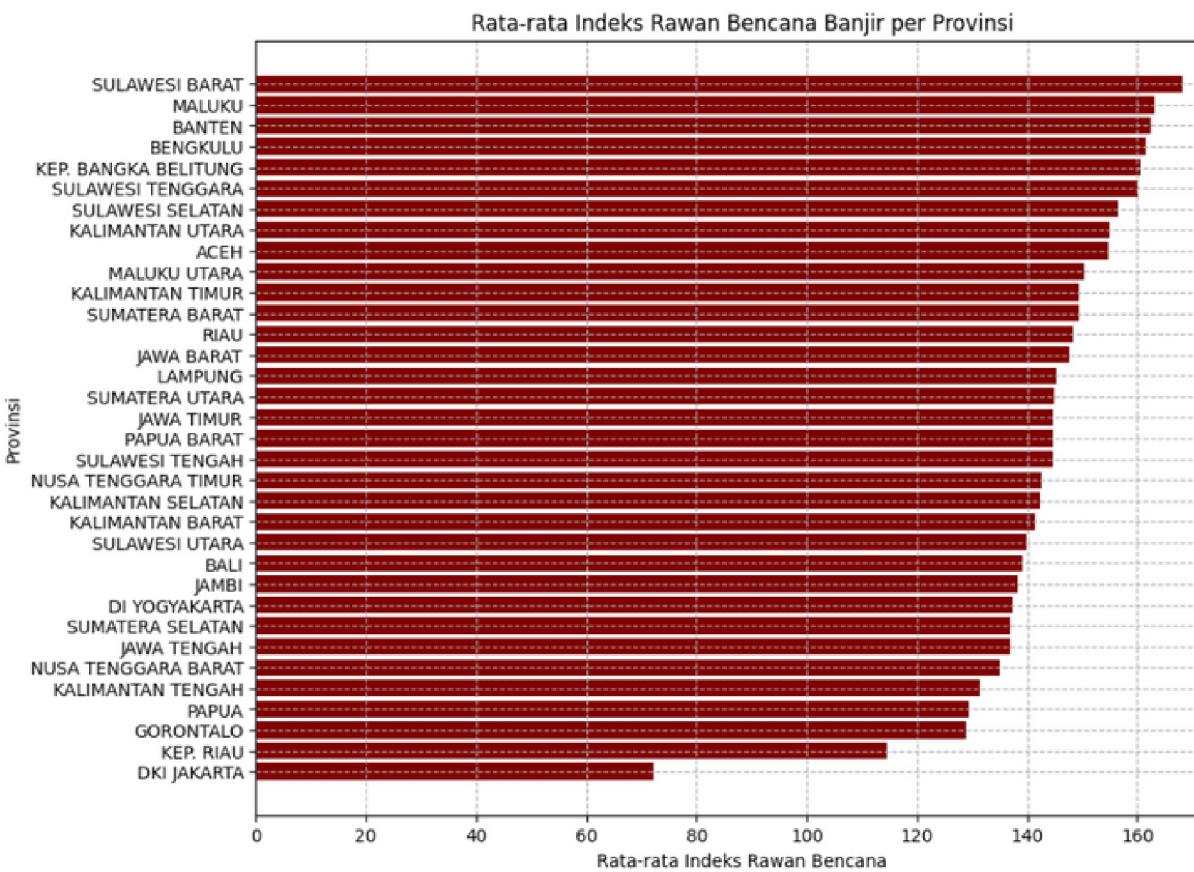
# Exploratory Data Analysis

Rerata IRBI  
Banjir per  
Provinsi



Sulawesi Barat

DKI Jakarta



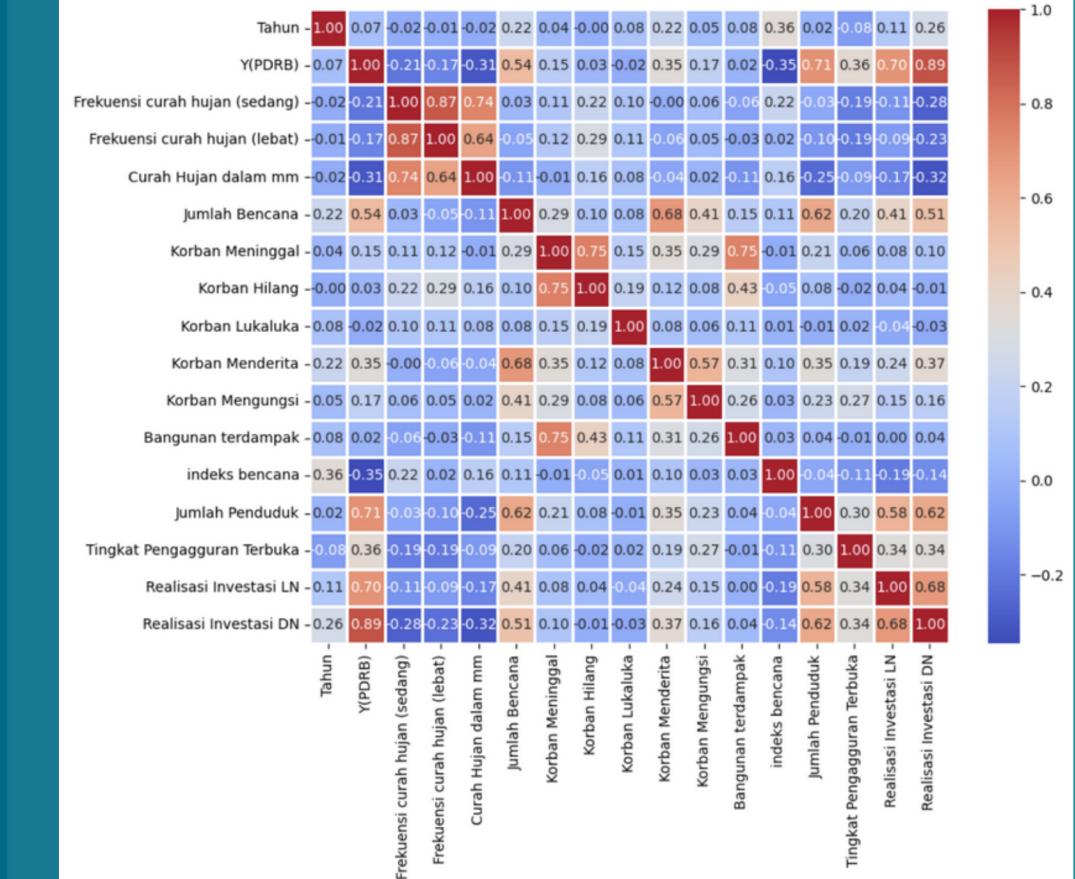
Korelasi  
antar  
Variabel

0,89

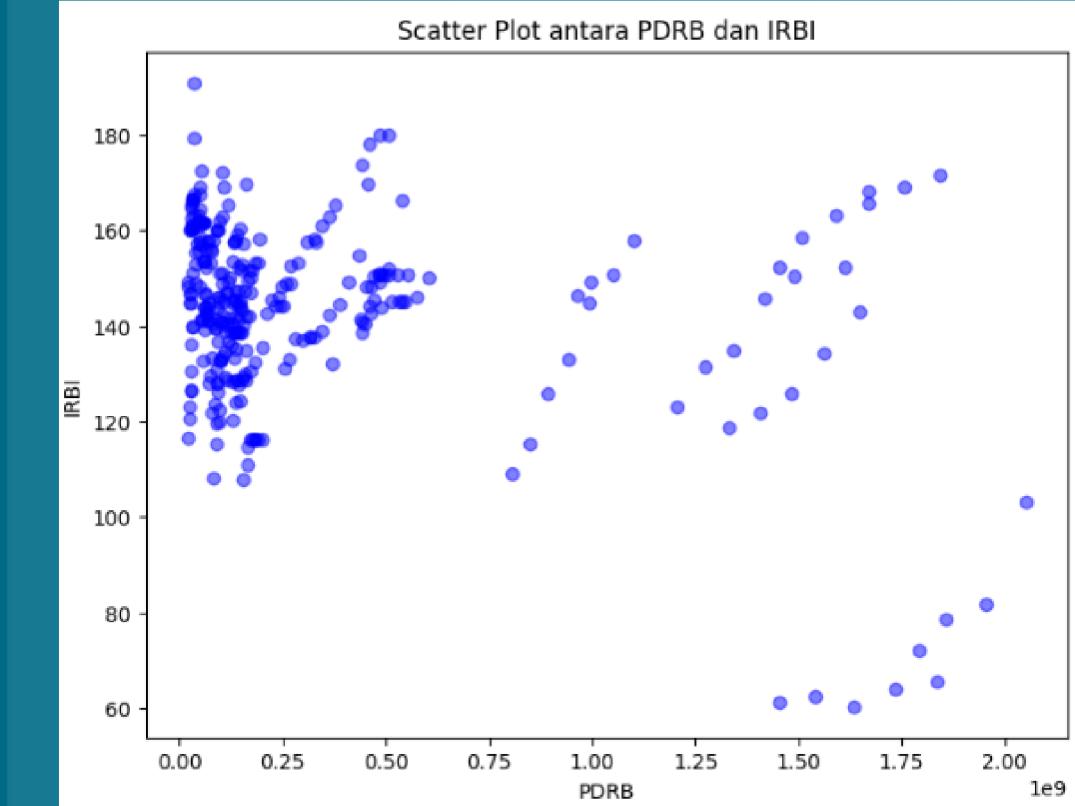
PDRB dengan  
Realisasi Investasi  
Dalam Negeri

-0,35

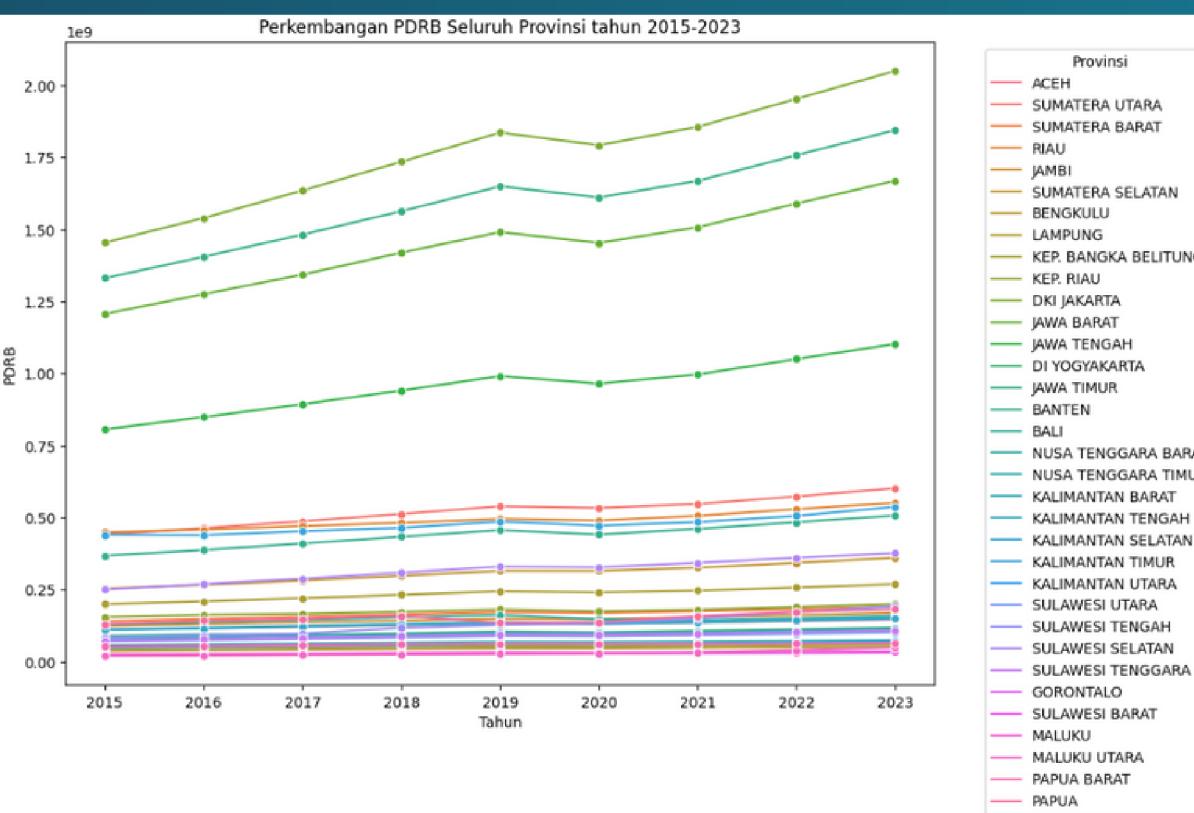
PDRB dengan  
Indeks Rawan  
Bencana Banjir



Hubungan  
antara PDRB  
dan IRBI



Indeks Rawan  
Bencana Banjir  
rendah  
cenderung  
memiliki PDRB  
tinggi



# Hasil Klasik

## Pengujian

### Uji Hausman

#### Hipotesis

$H_0$ : Model merupakan efek random

$H_1$ : Model merupakan efek tetap

#### Kesimpulan

Diperoleh nilai p – value =  $2.2e-16 < \alpha = 0.05$  maka  $H_0$  ditolak. Sehingga dapat disimpulkan bahwa model mengandung **efek tetap**.

### Diagnostic Checking

#### Normalitas

##### Hipotesis

$H_0$ : Galat berdistribusi normal

$H_1$ : Galat tidak berdistribusi normal

##### Kesimpulan

Diperoleh nilai p – value =  $0,002192 < \alpha = 0.05$  maka  $H_0$  ditolak. Sehingga dapat disimpulkan **galat tidak berdistribusi normal**

### Uji Breusch-Pagan

#### Hipotesis Efek individu

$H_0 : C = 0, d_t \sim iid, N(0, \sigma_d^2)$  (Tidak terdapat efek individu)

$H_1 : C \neq 0, d_t \sim iid, N(0, \sigma_d^2)$  (Terdapat efek individu)

#### Efek Waktu

$H_0 : D = 0, c_i \sim iid, N(0, \sigma_d^2)$  (Tidak terdapat efek waktu)

$H_1 : D \neq 0, c_i \sim iid, N(0, \sigma_d^2)$  (Terdapat efek waktu)

#### Efek Individu & Waktu

$H_0 : C = 0, D = 0$  (Tidak terdapat efek dua arah)

$H_1 : C \neq 0; D \neq 0$  (Terdapat efek dua arah)

#### Kesimpulan

Hipotesis	P-Value	Kesimpulan	Kesimpulan Akhir
Individu	< 2.2e-16	Ada efek individu	Hanya terdapat efek individu
Waktu	0.6920	Tidak ada efek waktu	
Individu & Waktu	< 2.2e-16	Ada efek dua arah	

### Uji Wald

#### Hipotesis Simultan

$H_0 : \text{Semua } \beta_i = 0 \text{ untuk } i = 1, 2, \dots, 15$

(model secara simultan tidak layak digunakan)

$H_1 : \text{Terdapat minimal satu } \beta_i \neq 0 \text{ untuk } i = 1, 2, \dots, 15$

(model secara simultan signifikan layak digunakan)

#### Parsial

$H_0 : \beta_i = 0 \text{ untuk } i = 1, 2, \dots, 15$

(Koefisien variabel independen tidak signifikan untuk masuk model)

$H_1 : \beta_i \neq 0 \text{ untuk } i = 1, 2, \dots, 15$

(Koefisien variabel independen signifikan untuk masuk model)

#### Kesimpulan

- Diperoleh nilai p – value =  $2.2e-16 < \alpha = 0.05$  maka  $H_0$  ditolak. Sehingga dapat disimpulkan bahwa **model layak digunakan**
- Realisasi Investasi Penanaman Modal Dalam Negeri Indeks Bencana, dan Jumlah Penduduk Signifikan.**



#### No Autocorrelation

##### Hipotesis

$H_0$ : Tidak ada korelasi serial galat

$H_1$ : Ada korelasi serial pada komponen galat

##### Kesimpulan

Diperoleh nilai p – value =  $0,002192 < \alpha = 0.05$  maka  $H_0$  ditolak. Sehingga dapat disimpulkan **ada korelasi serial pada komponen galat**

#### Homoskedastisitas

##### Hipotesis

$H_0$ : Homoskedastik

$H_1$ : Heteroskedastik

##### Kesimpulan

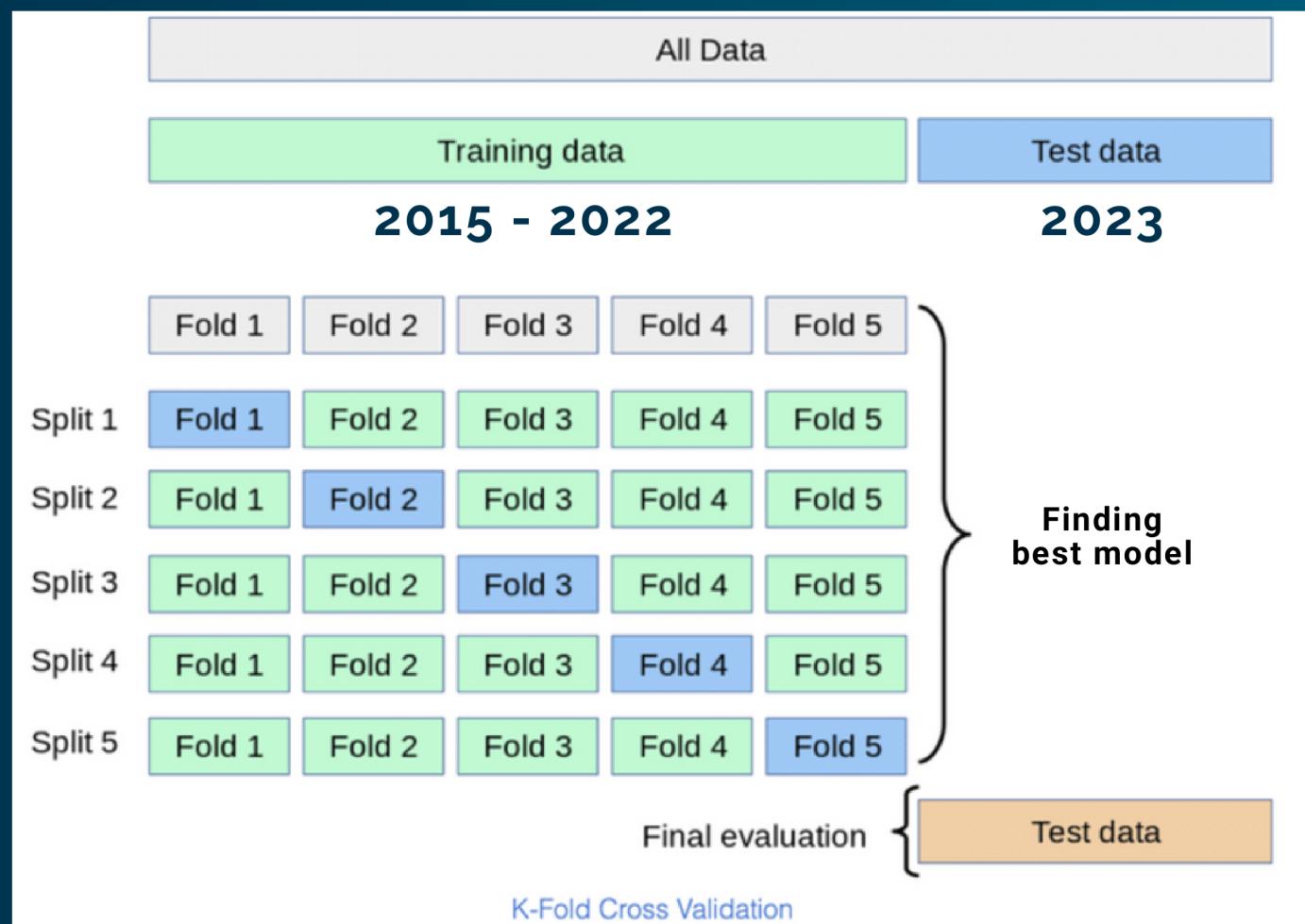
Diperoleh nilai p – value =  $0,002192 < \alpha = 0.05$  maka  $H_0$  ditolak. Sehingga dapat disimpulkan **galat bersifat heteroskedastik**

#### Kesimpulan

Model klasik analisis regresi linear data panel yang didapat adalah **fixed effect one-way dengan efek individu tanpa pengaruh waktu**. Namun pada analisis lanjutan pada tahap diagnostic checking model tersebut **tidak memenuhi asumsi normalitas, tidak ada korelasi serial, dan homoskedastisitas residual**. Sehingga **model klasik tidak dapat digunakan** untuk pemodelan

# Hasil Machine Learning

## Cross Validation



## K-Fold Cross Validation dengan k = 5 terhadap data train



model	RMSE	MAE	EV	R-square
XGBRegressor	1.926284e+08	7.485484e+07	8.296352e-01	0.827479
LinearRegression	1.618418e+08	1.102529e+08	8.715603e-01	0.866124
ElasticNet	3.735082e+08	2.621791e+08	2.765083e-01	0.260712
BayesianRidge	4.388316e+08	3.000155e+08	1.374456e-14	-0.028255
GradientBoostingRegressor	1.401976e+08	6.757467e+07	9.095366e-01	0.908169
CatBoostRegressor	1.483510e+08	7.012480e+07	8.986391e-01	0.896903
LGBMRegressor	1.587687e+08	9.473549e+07	8.759354e-01	0.871149

Model terbaik : **Gradient Boosting Regressor**

### Performa pada CV

- RMSE :  $1.401976 \cdot 10^8$
- MAE :  $6.757467 \cdot 10^7$
- EV : 0.9095366
- R-square : **0.908169**

### Performa pada data test

- RMSE :  $1.856451 \cdot 10^8$
- MAE :  $1.244616 \cdot 10^8$
- EV : 0.881744
- R-square : **0.86942**

# Hasil Machine Learning

## Hyperparameter Tuning Model Terbaik

## Random Search Cross Validation dengan n\_iter\_search = 40

### Kandidat hyperparameter

- n\_estimators : [100,200,300,400,500]
- learning rate : [0.01, 0.05, 0.1, 0.2]
- max\_depth : [3, 4, 5, 6, 7]
- min\_samples\_split : [2, 4, 6]
- min\_samples\_leaf : [1, 2, 3]

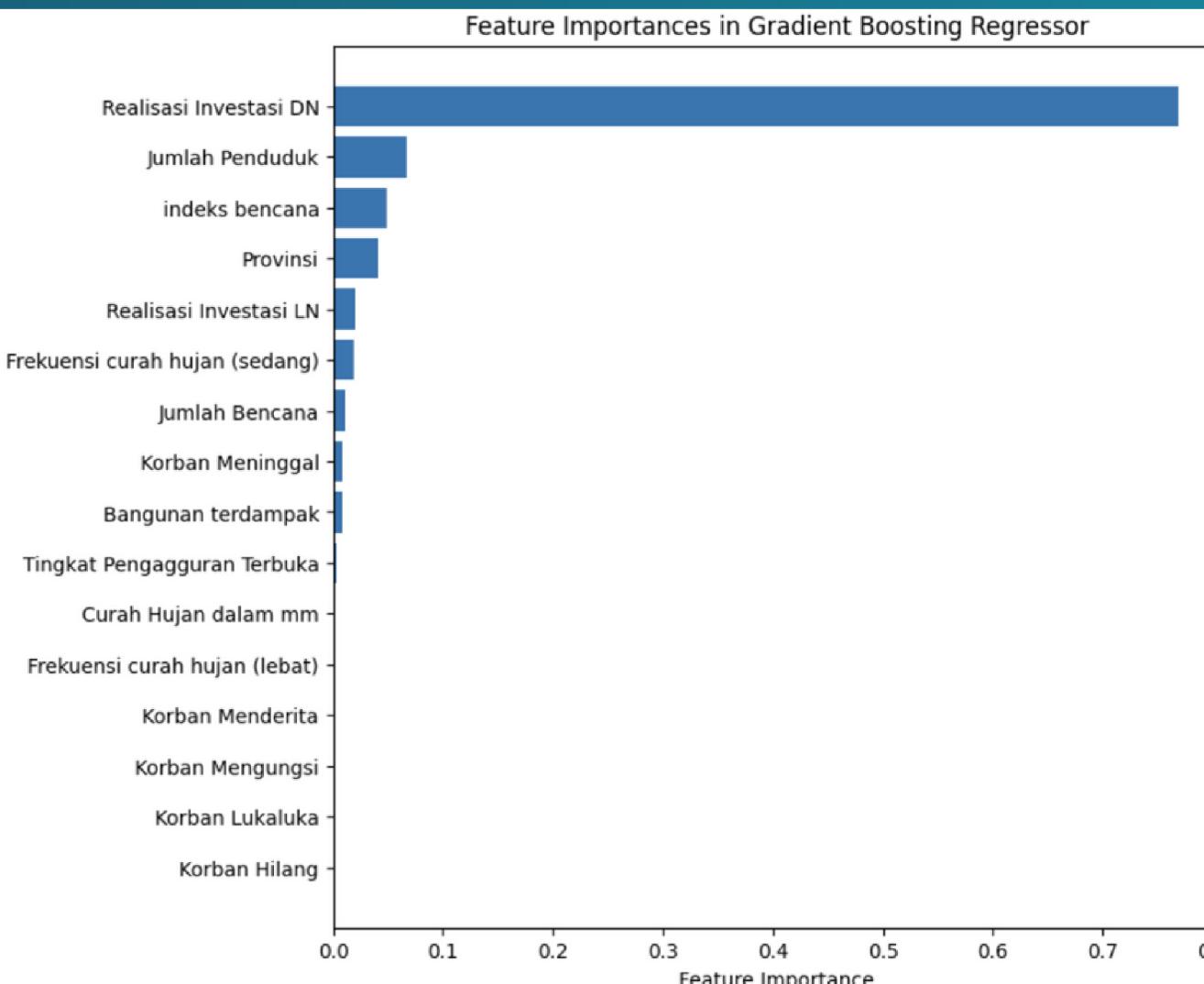


Random Search  
Cross Validation

### Parameter terbaik

- n\_estimators : 400
- learning\_rate : 0.1
- max\_depth : 3
- min\_samples\_split : 2
- min\_samples\_leaf: 1

### Feature Importance



### Variabel bencana

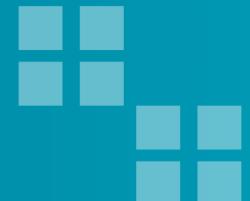
- Indeks Bencana (3)
- Frekuensi Curah Hujan Sedang (6)
- Jumlah Bencana (7)

### Variabel akibat bencana

- Korban meninggal (8)

### Performa pada RandomSearchCV

R-square : 0.8388128



### Performa pada data test

- RMSE :  $1.817550 \cdot 10^8$
- MAE :  $1.222458 \cdot 10^8$
- EV : 0.885653
- R-square : 0.874835



# Prediksi

**Best Model**

**Gradient Boosting Regressor**

**Hyperparameter:**

- 'n\_estimators': 400,
- 'min\_samples\_split': 2,
- 'min\_samples\_leaf': 1,
- 'max\_depth': 3,
- 'learning\_rate': 0.1

**Goodness of fit**

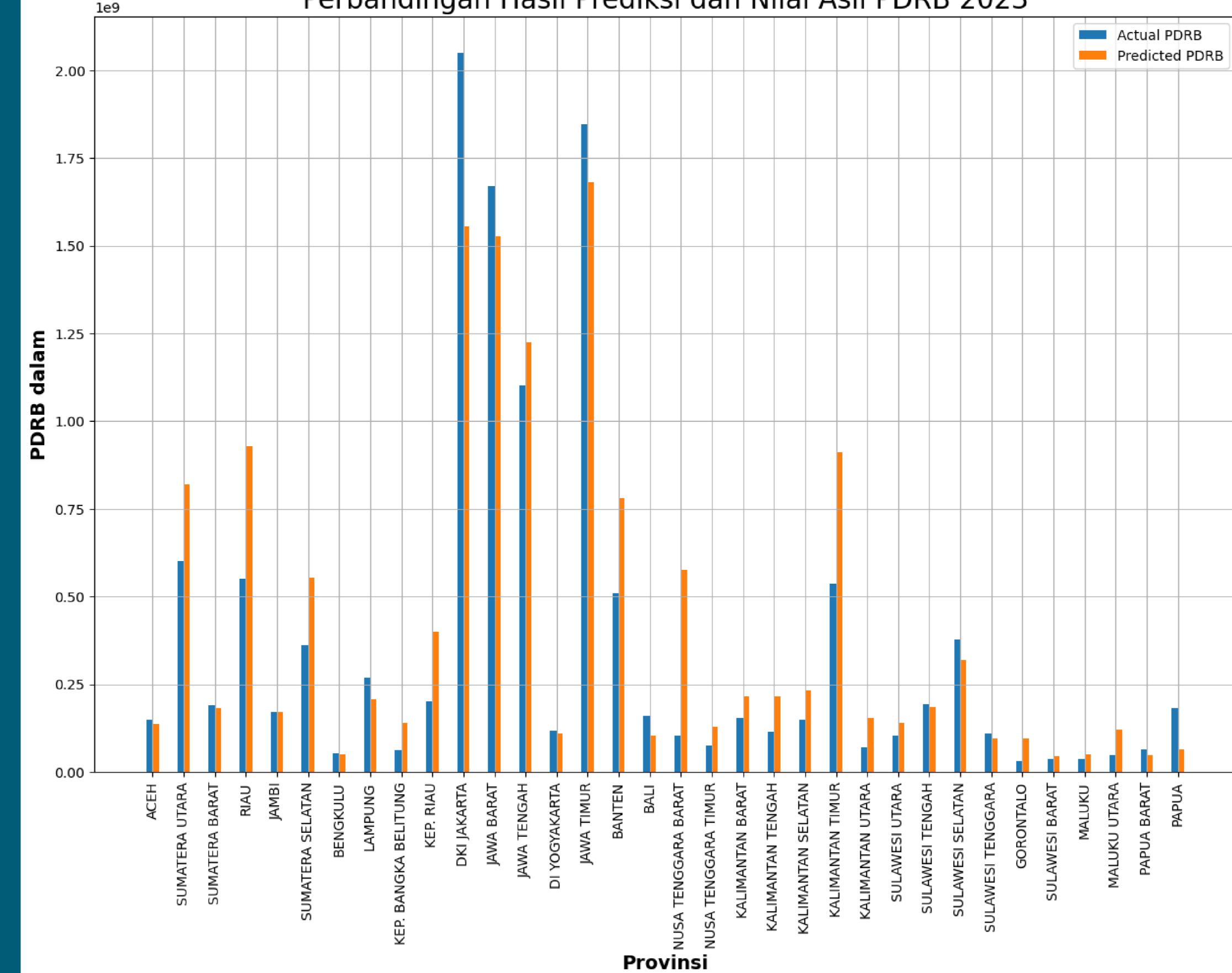
R square

**87 %**



**Prediksi 2023**

Perbandingan Hasil Prediksi dan Nilai Asli PDRB 2023



# Kesimpulan & Saran

## Kesimpulan

- **Gradient Boosting Regressor** terbukti paling efektif dengan meningkatkan prediksi dan mengurangi overfitting. Model ini berhasil memprediksi dampak bencana terhadap PDRB di 34 provinsi Indonesia (2015-2022) dengan goodness-of-fit **84 persen** dan nilai **87 persen** pada data testing 2023.

**Goodness of fit**

**Train** R square  
**84 %**      **Test** R square  
**87 %**

- Indikator paling berpengaruh adalah variabel non-bencana, yaitu realisasi **investasi modal dalam negeri, jumlah penduduk, dan provinsi**. Pada variabel bencana, indikator utama adalah **indeks bencana, frekuensi curah hujan sedang, jumlah bencana, korban meninggal, dan bangunan terdampak**



## Saran

- Model ini memiliki kemampuan generalisasi yang kuat, sehingga **peningkatan kualitas data bencana, terutama indeks risiko bencana dari BNPB**, sangat dianjurkan. Perbaikan ini termasuk peningkatan detail data dari tingkat provinsi ke kabupaten/kota.
- Selain itu, **melanjutkan dan memperbaiki basis data meteorologi** sangat penting juga untuk meningkatkan akurasi prediksi.
- Para pembuat kebijakan harus memberikan perhatian lebih pada daerah dengan **indeks bencana banjir** dan **frekuensi curah hujan tinggi** karena rentan terhadap banjir.



# TERIMAKASIH.

