PROYEK UJIAN AKHIR SEMESTER PENGANTAR DATA MINING

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

Tim ADAMANTINE

Anggota :

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

PROGRAM STUDI STATISTIKA

DEPARTEMEN MATEMATIKA

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS GADJAH MADA

2024

ABSTRAK

Sejalan dengan perubahan iklim yang semakin meluas, cepat, dan intens, frekuensi curah hujan lebat dan banjir terus meningkat. Makalah ini menganalisis pengaruh bencana banjir dan kondisi ekonomi regional terhadap Produk Domestik Regional Bruto (PDRB) di Indonesia menggunakan teknik *machine learning* yang bertujuan untuk menyelesaikan masalah regresi. Pemodelan dengan metode klasik menggunakan Regresi Data Panel juga dilakukan sebagai perbandingan. Model prediksi dibangun menggunakan kumpulan fitur dari berbagai dimensi data, termasuk fitur bencana seperti curah hujan, frekuensi curah hujan sedang, frekuensi curah hujan lebat, dan jumlah bencana. Selain itu, fitur akibat bencana yang digunakan meliputi jumlah korban meninggal, korban hilang, korban luka-luka, korban menderita, korban mengungsi, bangunan terdampak, dan indeks bencana banjir. Variabel ekonomi yang digunakan meliputi provinsi, jumlah penduduk, tingkat pengangguran terbuka, realisasi investasi penanaman modal luar negeri, dan realisasi investasi penanaman modal dalam negeri. Berdasarkan data historis banjir dari 34 provinsi di Indonesia pada periode 2015-2023, metode *machine learning* dan metode klasik dibandingkan untuk memprediksi dampak banjir terhadap PDRB. Hasil analisis menunjukkan bahwa model *machine learning Gradient Boosting Regression* memiliki kinerja terbaik dengan nilai *R-square* sebesar 87,5%. Pemodelan menunjukkan bahwa fitur bencana seperti indeks bencana, frekuensi curah hujan sedang, dan jumlah bencana banjir termasuk dalam kategori *feature importance* yang tinggi. Hal ini mengindikasikan bahwa bencana banjir memiliki dampak signifikan terhadap PDRB di provinsi-provinsi Indonesia.

Kata Kunci: Regresi, *Machine Learning*, Data Panel, Banjir, PDRB

PENDAHULUAN

A. Latar Belakang

Tidak dapat dipungkiri bahwa Indonesia adalah daerah yang sangat rawan terhadap bencana alam. Menurut *World Risk Report*, Indonesia berada di peringkat kedua negara yang paling berisiko bencana di dunia, dengan indeks risiko bencana sebesar 43,5 dari 100 (WRI, 2022). Salah satu bencana alam yang sering terjadi di Indonesia adalah banjir, yang pada tahun 2022 tercatat sebagai bencana yang paling sering terjadi, dengan jumlah kejadian sebanyak 1.520.

Bencana banjir yang kerap terjadi di Indonesia secara langsung berdampak pada kehidupan sehari-hari masyarakat, infrastruktur, bahkan aset negara. Menurut Badan Nasional Penanggulangan Bencana (BNPB), jumlah jiwa terpapar risiko banjir tersebar di beberapa pulau di Indonesia, dengan jumlah melebihi 170 juta jiwa dan nilai aset terpapar melebihi Rp 750 triliun (BNPB, 2022). Dalam skala yang lebih kecil misalkan pada skala provinsi, bencana banjir memberikan dampak yang besar terhadap sektor perekonomian suatu daerah. Sebagai contoh kerugian ekonomi akibat banjir Jakarta tahun 2013 mencapai Rp 1 triliun. Banjir ini menyebabkan tidak berjalannya aktivitas perdagangan. Tidak hanya itu, gardu listrik yang terendam banjir mengakibatkan kawasan industri lumpuh karena tak memperoleh suplai listrik. Kejadian tadi adalah contoh angka kerugian dan dampak dari banjir pada satu waktu dan satu titik. Indonesia merupakan negara dengan 34 provinsi yang rawan terhadap bencana terutama banjir. Secara keseluruhan, bentuk kerugian dan dampak ini tidak dapat dianggap remeh.

Kondisi ekonomi di suatu daerah dapat dilihat menggunakan beberapa indikator ekonomi. Salah satunya adalah Produk Domestik Regional Bruto (PDRB) yang menjelaskan kemampuan suatu daerah dalam mengelola perekonomiannya. Salah satu pendekatan perhitungan PDRB adalah pedekatan atas dasar harga konstan menurut pengeluaran. Dengan begitu secara intuitif terdapat praduga bahwa kejadian banjir akan memengaruhi PDRB pada suatu provinsi.

B. Tujuan dan Manfaat

Tujuan dari penelitian ini adalah sebagai berikut:

- Melihat hubungan antara PDRB provinsi dengan kejadian dan akibat bencana banjir.
- Membentuk model regresi yang baik dalam memprediksi PDRB pada setiap provinsi di Indonesia bedasarkan variabel-variabel yang menjelaskan mengenai kejadian banjir, akibat banjir, serta indikator ekonomi daerah.
- Mengidentifikasi faktor-faktor apa saja yang berpengaruh secara signifikan terhadap PDRB suatu provinsi.

Melalui penelitian ini, model prediksi yang akan dibentuk dapat digunakan sebagai kajian tambahan bagi pembuat kebijakan untuk meminimalisasi dampak ekonomi yang diakibatkan oleh bencana banjir.

METODE ANALISIS DATA

A. Data

Penelitian ini menerapkan *data integration* untuk mengumpulkan data yang dibutuhkan dari berbagai sumber yang berbeda. (Cuello, 2023). Metode ini menggabungkan data dari banyak sumber menjadi satu data terpadu yang konsisten, akurat, dan dapat digunakan untuk analisis. Proses ini melibatkan beberapa teknik untuk mengatasi ketidakkonsistenan data dan perbedaan format. Data yang akan digunakan pada penelitian ini bersumber dari empat situs utama, yaitu situs Badan Nasional Penanggulangan Bencana (BNPB), Indeks Risiko Bencana Indonesia (inaRISK), Badan Pusat Statistik (BPS), dan World Bank (WorldBank, 2021).

Terdapat dua tahap utama yang dilakukan pada proses ini, yaitu tahap ekstraksi data dari setiap sumber dan tahap penggabungan data. Tahap ekstraksi mengambil informasi yang diperlukan dalam setiap situs. Untuk situs BNPB dan inaRISK, tahap ekstraksi data dilakukan menggunakan metode *web scrapping* dengan `selenium` (David, n.d.). Metode ini melakukan pengambilan data dari situs menggunakan otomatisasi yang mengontrol *browser* situs dinamis. `Selenium` melakukan simulasi interaksi dengan situs, yaitu seperti menavigasi halaman *website*, memilih konten atau tombol yang spesifik, dan lain sebagainya. Menggunakan metode ini, data indeks risiko bencana banjir serta korban dan infrastruktur terdampak akibat bencana banjir di 34 provinsi pada tahun 2015 sampai 2023. Sedangkan pada situs BPS dan World Bank, data mengenai angka PDRB, jumlah penduduk, tingkat pengangguran terbuka, realisasi investasi dalam dan luar negeri, curah hujan, serta frekuensi hujan sedang dan lebat telah tersedia untuk ke-34 provinsi tahun 2015 sampai 2023. Tahap *data integration* selanjutnya adalah menggabungkan data menjadi terstruktur dan terpadu. Data-data yang telah diekstrak dari berbagai sumber digabungkan menurut nama provinsi dan tahunnya.

Data yang terintegrasi menghasilkan 306 observasi dari ke-34 provinsi Indonesia dalam 9 tahun (2015-2023). Terdapat 16 variabel prediktor dimana satu variabel merupakan variabel kategorik dan sisanya merupakan variabel numerik.

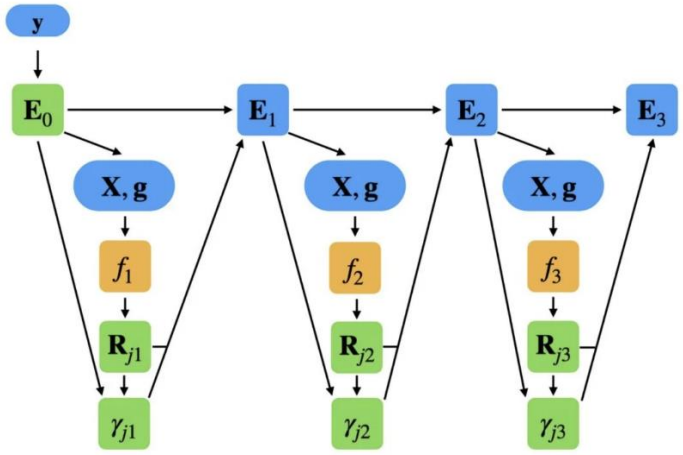
Variabel-variabel ini dikelompokkan menjadi tiga kelompok yang dirincikan sebagai berikut:

Kelompok	***Nama Variabel***	***Keterangan***
Variabel Bencana	Frekuensi curah hujan sedang	Rata-rata frekuensi kejadian saat curah hujan harian mencapai 20 mm
	Frekuensi curah hujan lebat	Rata-rata frekuensi kejadian saat curah hujan harian mencapai 50 mm
	Jumlah bencana	Frekuensi kejadian bencana banjir tahunan
	Curah hujan	Rata-rata curah hujan tahunan dalam mm
Variabel Akibat Bencana	Korban meninggal	Jumlah korban meninggal akibat banjir per tahun
	Korban hilang	Jumlah korban hilang akibat banjir per tahun
	Korban luka-luka	Jumlah korban luka-luka akibat banjir per tahun
	Korban menderita	Jumlah korban menderita akibat banjir per tahun
	Korban mengungsi	Jumlah korban mengungsi akibat banjir per tahun
	Bangunan terdampak	Jumlah bangunan terdampak akibat banjir per tahun
	Indeks bencana banjir	Indeks rawan bencana banjir serkisar
Variabel Ekonomi	Provinsi	Nama provinsi
	Jumlah penduduk	Jumlah penduduk tahunan
	Tingkat pengangguran terbuka	Tingkat pengagguran terbuka dalam persen
	Investasi Luar Negeri	Realisasi investasi penanaman modal luar negeri dalam Juta US$
	Investasi Dalam Negeri	Realisasi investasi penanaman modal dalam negeri dalam Juta US$
Variabel Target	PDRB	PDRB atas dasar harga konstan menurut pengeluaran dalam juta Rupiah

B. Metode Pemodelan

1. *Machine Learning* dengan *Gradient Boosting Regressor*

Gradient Boosting Regressor adalah salah satu model keluarga *gradient boosting machine* yang membangun model secara bertahap dan menggabungkannya (*ensemble*) untuk meningkatkan performa model dalam menyelesaikan masalah regresi. *Gradient boosting machines*, atau yang biasa disebut GBMs, merupakan salah satu prosedur pembelajaran mesin yang secara berturut-turut menyesuaikan model baru untuk memberikan perkiraan yang lebih akurat terhadap variabel respons. Prinsip utama di balik *GBM* ini adalah membangun *base-learner* baru agar seoptimal mungkin berkorelasi dengan gradien negatif dari *loss function* yang terkait dengan seluruh *ensemble*. *Loss function* yang diterapkan bisa bermacam-macam. Secara umum, pemilihan *loss function* tidak memiliki aturan khusus dengan banyak variasi *loss function* yang telah dikembangkan sejauh ini dan dengan kemungkinan mengimplementasikan *loss function* sesuai dengan tujuan tertentu.



Gambar 1. *Ilustrasi Algoritma Gradient Boosting Regressor*

Algoritma *Gradient Boosting Regressor* secara umum memiliki alur yang sama dengan algoritma GBM (Datamapu, 2024). Ilustrasi algoritma *Gradient Boosting Regressor* ditunjukkan pada Gambar 1. Secara ringkas algoritma ini dapat dijelaskan menjadi seperti berikut:

Input

Data train $(x_i, y_i)_{i=1}^{n}$, loss function $L(y, E(x))$, dan jumlah iterasi (*weak learners*) M.

Step 1 **:** Inisialisasi model dengan *single leaf* yang bernilai konstan.

$$E_0(x) = arg\ min_\gamma \sum_{i=1}^{n} L(y_i, \gamma)$$

Step 2 **:** Membuat *weak learners* sebanyak M dalam *loop.*

1.) Menghitung residual dengan mendiferensiasikan *loss function* untuk mendapatkan *gradient* (g).

$$R_{im} = -\left[\frac{\delta\, L(y_i, E(x_i)}{\delta\, E(x_i)}\right]_{E(x)=E_{m-1}(x)} \qquad for\ i = 1, \dots, n$$

2.) Membangun *weak learner* $f_m(x)$ berupa decision tree dengan variabel target adalah residual R_m.

3.) Menghitung nilai γ yang meminimumkan *loss function*

$$\gamma_m = arg\ min_\gamma \sum_{i=1}^{n} L\big(y_i, E_{m-1}(x_i)\big) + \gamma\, f_m(x_i)$$

4.) Meng-*update* model dengan mempertimbangkan *learning rate*.

$$E_m(x) = E_{m-1}(x) + \gamma\, f_m(x)$$

Step 3 **:** Model akhir $E_m(x)$ terbentuk.

Model *Gradient Boosting Regressor* memiliki kelebihan dalam kemampuannya dalam mengatasi *missing value* dan *outlier* pada variabel prediktor, fleksibilitas tipe data input, dan multifungsional dikarenakan variabilitas *loss function* yang dapat digunakan. Namun model ini juga memiliki kekurangan yaitu sensitifitasnya terhadap *outlier* di variabel target dan pemilihan jumlah *weak learners* yang mungkin menyebabkan adanya *overfit*.

2. Analisis Regresi Linear Data Panel

Gabungan antara data runtun waktu dan data *cross section* disebut data panel, yaitu data yang dikumpulkan dalam kurun waktu tertentu untuk sejumlah objek atau lokasi tertentu (Rosadi, 2022). Teknik analisis yang mencoba menjelaskan bentuk hubungan antara peubah-peubah yang mendukung sebab-akibat menggunakan data panel disebut model regresi data panel. Terdapat 3 macam estimasi yang umum dilakukan yaitu Fixed Effect Model dan Random Effect Model.

2.1. Persamaan umum Regresi Data Panel

Estimasi regresi data panel yang menggabungkan seluruh data baik cross section ataupun time series dengan tidak menghiraukan efek waktu dan lokasi. Nila intercept diasumsikan sama untuk masing asing fitur begitu jga untuk koefisien pada semua unit *cross section* dan *time series*. Persamaan *Common Effect Model* dinyatakan dalam:

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

Keterangan:

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

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

β_k : Koefisien variabel independen ke-k, $k = 1, 2, ...$

C_i : Konstanta yang bergantung pada individu ke-i

d_t : Konstanta yang bergantung pada waktu ke-t

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

2.2. *Fixed Effect Model*

Fixed Effect Model adalah metode regresi yang mengestimasi data panel dengan menambahkan variabel boneka (*dummy*). Model ini mengasumsi bahwa terdapat efek yang berbeda antar individu. Perbedaan itu dapat diakomodasi melalui perbedaan pada intersepnya. Oleh karena itu, dalam

Fixed Effect Model setiap individu merupakan parameter yang tidak diketahui dan akan diestimasi dengan menggunakan teknik *Least Square Dummt Variable*. Persamaan *Fixed Effect* yaitu

$$y_{i,t} = x'_{i,t}\beta + c_i + d_t + \varepsilon_{i,t}$$

Di mana c_i merupakan konstanta yang bergantung pada unit ke-i, tetapi tidak pada waktu t. d_t merupakan konstanta yang bergantung pada waktu t, tetapi tidak pada unit i.

Di sini apabila memuat komponen c_i dan d_t, model disebut model efek tetap dua arah, sedangkan apabila $d_t = 0$ atau $c_i = 0$, model disebut model efek tetap satu arah.

2.3. *Random Effect Model*

Dengan menggunakan model efek tetap, kita tidak dapat melihat pengaruh dari berbagai karakteristik yang bersifat konstan dalam waktu atau konstan di antara individual. Untuk maksud tersebut kita dapat menggunakan model efek acak (*random effect*) yang secara umum dituliskan sebagai

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

Di mana $v_{i,t} = C_i + d_t + \varepsilon_{i,t}$. Di sini c_i diasumsikan bersifat *independent and identically distributed* (i.i.d.) normal dengan mean 0 dan variansi σ_c^2, d_t diasumsikan bersifat i.i.d. normal dengan mean 0 dan variansi σ_d^2, dan $\varepsilon_{i,t}$ bersifat i.i.d. normal dengan mean 0 dan variansi σ_c^2 (dan $\varepsilon_{i,t}$, c_i dan d_t diasumsikan independen satu dengan lainnya). Jika komponen d_t atau c_i diasumsikan 0, model disebut model efek acak satu arah, sedangkan pada keadaan lain disebut model dua arah. Terdapat tiga jenis uji khusus yang digunakan untuk memilih model regresi data panel yang terbaik untuk suatu permasalahan yang ada, yaitu uji chow, uji hausman, dan uji lagrange multiplier.

2.4. Uji Hausman

Uji Hausman bertujuan untuk memilih antara model FEM dan model REM. Dengan mengikuti kriteria Wald, nilai statistik Hausman dapat dihitung dengan rumusan sebagai berikut

$$W = \chi^2_{(p)} = [b - \beta]' \psi^{-1} [b - \beta]$$
$$\psi = Var[b] - var[\beta]$$

Di mana β adalah parameter (tanpa intersept) random effect dan β adalah parameter *fixed effect* menggunakan LSDV. Var[b] merupakan matriks kovarian parameter (tanpa intersep) *random effect* dan var[β] adalah matriks kovarian parameter *fixed effect*. Apabila nilai $W > \chi^2_{(\alpha,p)}$, maka model yang terpilih adalah model FEM. P adalah jumlah variabel independen

2.5. Uji Breusch Pagan

Uji bertujuan untuk melihat apakah terdapat efek *cross-section time* (atau keduanya) di dalam panel data, yaitu dengan menjadi hipotesis berbentuk:

$H_0: c = 0, d = 0$ atau tidak terdapat efek *cross section* maupun *time*

$H_0: c = 0$ atau tidak terdapat efek *cross section*

$H_0: d = 0$ atau tidak terdapat efek waktu

Secara umum, langkah-langkah uji hipotesis yang dilakukan adalah sebagai berikut: pertama-tama dilakukan uji Hausman terhadap data jika hipotesis untuk uji hausman ditolak maka model *fixed effect* digunakan dalam pemodelan. Selanjutnya, dilakukan uji Breusch Pagan untuk melihat apakah terdapat efek waktu dan/atau *cross section* di dalam data. Jika hipotesis Breusch Pagan tidak ditolak maka dilakukan analisis dengan menggunakan model regresi panel (Rosadi, 2022). Setelah melakukan uji spesifikasi dan didapatkan model yang tepat dalam menggambarkan data maka dilakukan uji asumsi klasik yang terdiri dari beberapa asumsi yang harus dipenuhi antara lain uji Normalitas, Multikolinearitas, Heteroskedastik dan Autokorelasi. Dilanjutkan dengan uji Hipotesis yakni: Uji hipotesis terhadap

masing-masing koefisien regresi (Uji t), Uji hipotesis regresi secara menyeluruh (Uji F) dan Koefisien determinasi.

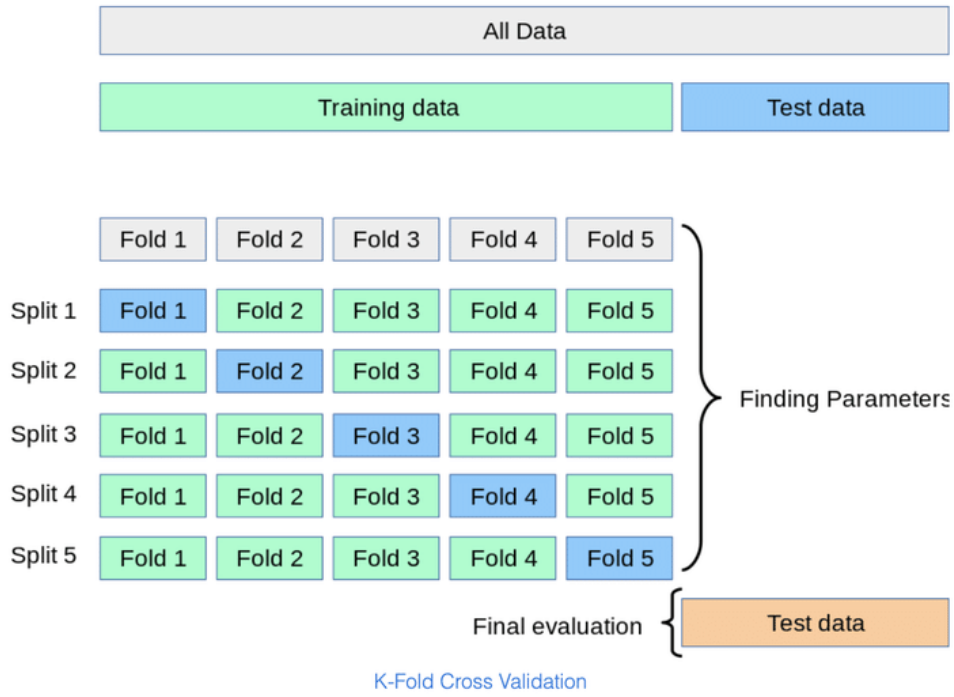
3. Metriks Evaluasi

Model prediksi yang dihasilkan dari regresi harus divalidasi dengan k-fold cv, yang merupakan salah satu metode CV yang direkomendasikan untuk memilih model terbaik karena cenderung memberikan estimasi akurasi yang tidak terlalu bias dibandingkan dengan jenis CV lainnya. *Cross-validation* (CV) merupakan salah satu teknik *machine learning* untuk memvalidasi model prediktif. Dalam prosesnya, dilakukan pembagian data menjadi dua bagian, yaitu satu bagian untuk melatih model (*training* data) dan bagian lainnya untuk menguji model (*testing* data). Salah satu metode CV yang umum digunakan adalah K-Folds, dimana data dibagi menjadi k-sub data. Seperti yang terdapat pada Gambar 1, salah satu sub data digunakan sebagai data pengujian dan sisanya digunakan sebagai data pelatihan. Proses ini diulangi sebanyak k-kali dengan setiap sub data digunakan sekali sebagai data pengujian.

Model prediksi yang dihasilkan dari regresi harus divalidasi menggunakan k-fold *cross-validation* (CV), yang merupakan salah satu metode CV yang direkomendasikan untuk memilih model terbaik. K-fold CV cenderung memberikan estimasi akurasi yang tidak terlalu bias dibandingkan dengan jenis CV lainnya. *Cross-validation* (CV) adalah teknik dalam *machine learning* untuk memvalidasi model prediktif dengan cara membagi data menjadi dua bagian: satu bagian untuk melatih model (*training* data) dan bagian lainnya untuk menguji model (*testing* data).

Salah satu metode CV yang umum digunakan adalah K-Folds. Dalam metode ini, data dibagi menjadi k sub-data (*folds*). Seperti yang ditunjukkan pada Gambar 2, salah satu sub-data digunakan sebagai data pengujian, sementara sisanya digunakan sebagai data pelatihan. Proses ini diulangi sebanyak k kali, dengan setiap sub-data digunakan sekali sebagai data pengujian. Dengan demikian, setiap observasi digunakan sekali sebagai data pengujian dan k-1 kali sebagai data

pelatihan. Metode ini memastikan bahwa model dievaluasi secara menyeluruh dan memberikan gambaran yang lebih akurat tentang performa model pada data yang belum pernah dilihat sebelumnya.



Gambar 2. *K-fold CV*

Dalam membandingkan perorma validasi dari semua model, 4 indikator digunakan, termasuk *Root-Mean-Square Error* (RMSE), *Mean Absolute Error* (MAE), *Explained Variance* (EV), dan *R-square score* (R^2). Formula yang digunakan sebagai berikut:

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

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

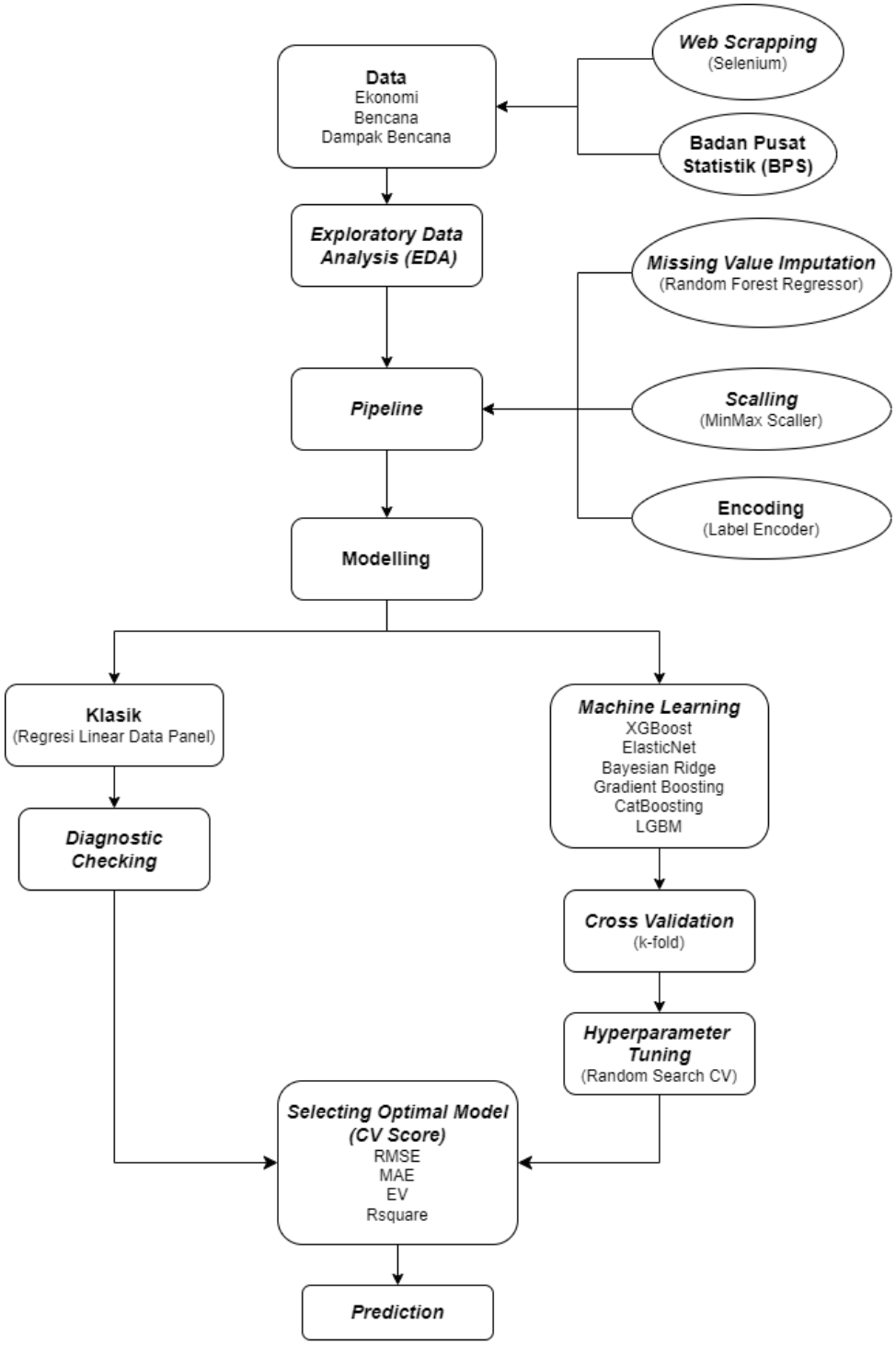
$$EV(y_i, \hat{y}_i) = 1 - \frac{Var(y_i - \hat{y}_i)}{Var(y_i)}$$

$$R^2(y_i, \hat{y}_i) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{RMSE}{Var(y_i)}$$

Di mana y_i adalah nilai observasi asli, $\hat{y}_i$ merupakan nilai prediksi, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, n adalah jumlah observasi. Pada setiap pengujian sebanyak k=5 akan dihasilkan metrik evaluasi kinerja model. Hasilnya kemudian dirata-ratakan untuk mendapatkan estimasi yang lebih baik tentang kinerja model. Mode terbaik dipilih berdasarkan yang paling banyak memenuhi kriteria RMSE dan MAE terkecil serta MAE dan EV terbesar.

4. Alur Penelitian

Keseluruhan proses penelitian ini ditunjukkan pada gambar 3



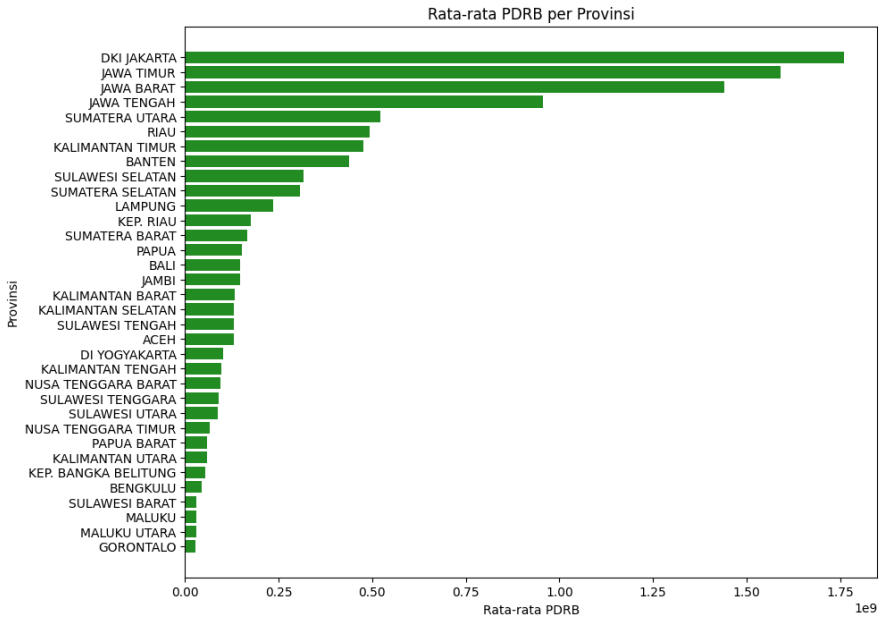
Gambar 3. *Diagram Alir Penelitian*

HASIL ANALISIS

A. *Exploratory Data Analysis*

Exploratory data analysis (EDA) adalah suatu pendekatan analisis yang bertujuan untuk mengidentifikasi pola dan karakteristik data secara umum. Biasanya, EDA dilakukan dengan menampilkan dan mengamati grafik atau visualisasi dari data. Dengan demikian, pengetahuan mengenai data mulai dari bentuk sebaran, outlier, proporsi kategori, hingga anomali akan dapat terdeteksi. Beberapa visualisasi telah dibuat sebagai langkah awal mencari tahu karakteristik dari data yang dimiliki. Berikut adalah visualisasi dan interpretasinya:

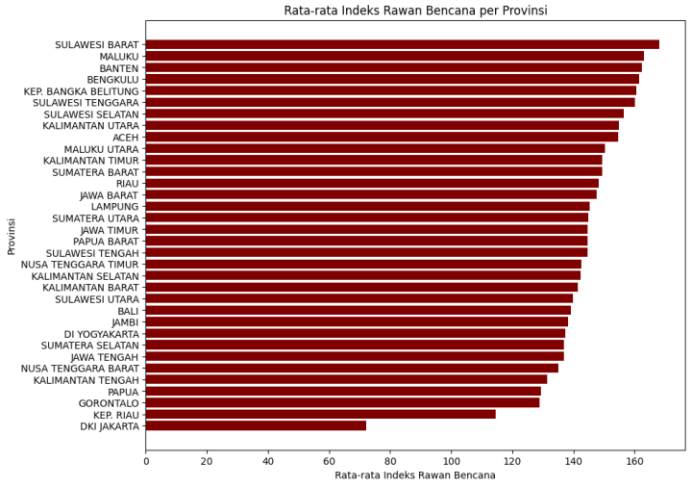
1. Rata-rata PDRB provinsi tahun 2015-2023



Gambar 4. *Rata-rata PDRB per provinsi tahun 2015-2023*

Barchart di atas menunjukkan rata-rata PDRB per provinsi dari tahun 2015-2023. Didapatkan tiga provinsi dengan rata-rata nilai PDRB tertinggi adalah DKI Jakarta, Jawa Timur, dan Jawa Barat. Sedangkan, tiga provinsi dengan rata-rata nilai PDRB terendah adalah Maluku, Maluku Utara, dan Gorontalo di peringkat akhir.

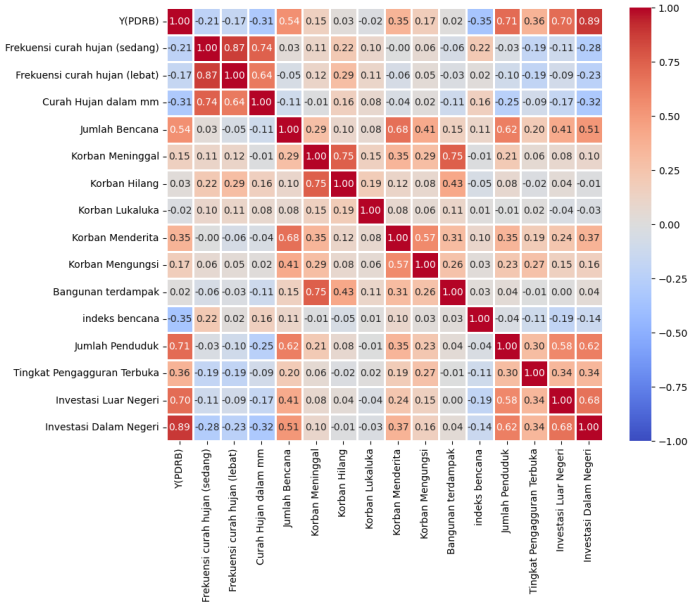
2. Rata-rata indeks rawan bencana banjir setiap provinsi tahun 2015-2023



Gambar 5. *Rata-rata Indeks Rawan Bencana Banjir per Proviinsi*

Barchart di atas menunjukkan rata-rata indeks rawan bencana banjir per provinsi dari tahun 2015-2023. Didapatkan tiga provinsi dengan rata-rata indeks rawan bencana banjir tertinggi adalah Sulawesi Barat, Maluku, dan Banten. Sedangkan, tiga provinsi dengan rata-rata indeks rawan bencana banjir terendah adalah Gorontalo, Kepulauan Riau, dan DKI Jakarta di peringkat akhir.

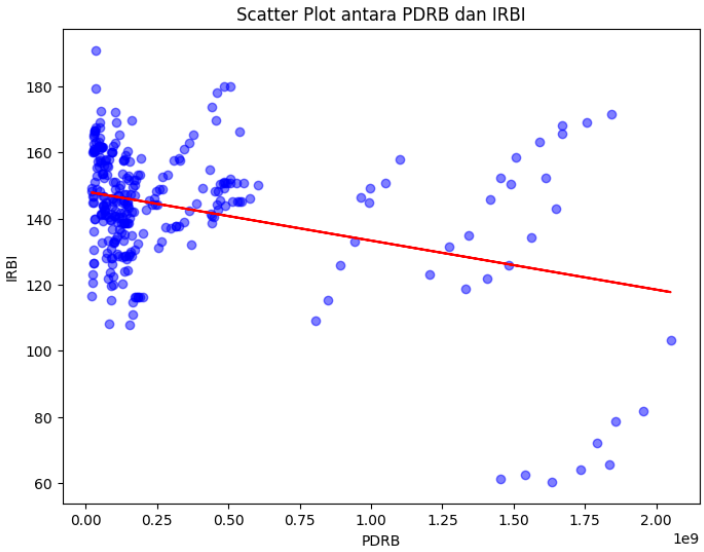
3. Korelasi antara tiap variabel numerik



Gambar 6. *Heatmap Matrix Correlation antar fitur numerik*

Heatmap Correlation Plot di atas menunjukkan korelasi dari setiap pasang variabel numerik. Kotak berwarna kemerahan menandakan adanya korelasi positif sedangkan kotak berwarna kebiruan menandakan adanya korelasi negatif. Warna yang semakin kontras baik berwarna merah atau biru menandakan korelasi yang kuat sedangkan warna yang tidak kontras atau keabuan menandakan tidak adanya korelasi. Didapatkan bahwa pasangan variabel yang memiliki hubungan terkuat secara positif adalah variabel PDRB dan Investasi Dalam Negeri dengan nilai koefisien korelasi sebesar 0,89. Pasang variabel yang memiliki hubungan terkuat secara negatif adalah PDRB dan Indeks Rawan Bencana Banjir dengan nilai koefisien korelasi sebesar -0,35.

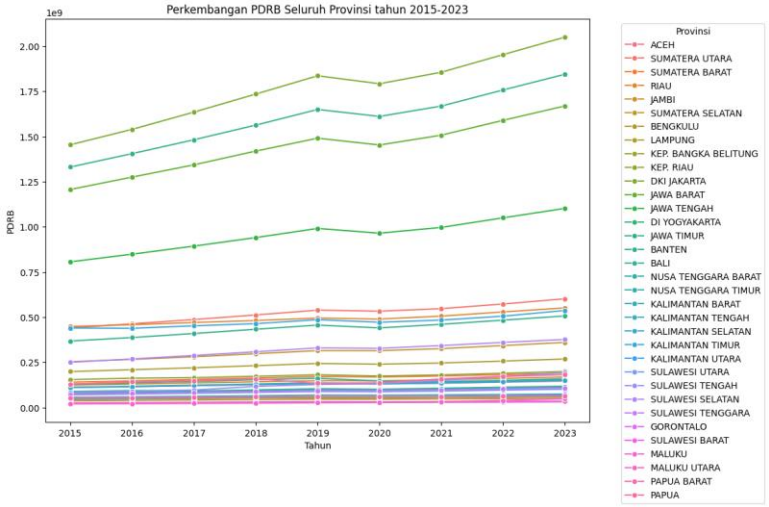
4. Hubungan antara Indeks Rawan Bencana dengan PDRB



Gambar 7. *Scatter plot antara variabel PDRB dengan indeks rawan bencana banjir*

Scatter plot di atas memberikan gambaran mengenai hubungan antara variabel PDRB dan indeks rawan bencana banjir. Didapatkan bahwa kedua variabel ini tidak memiliki hubungan linear yang signifikan. Namun, dapat ditarik kesimpulan bahwa apabila indeks rawan bencana banjir bernilai rendah maka PDRB cenderung bernilai tinggi.

5. Pola pergerakan nilai PDRB provinsi tiap tahunnya



Gambar 8. *Multiple Line Chart tren PDRB tiap provinsi*

Multiple Line Chart di atas menunjukkan pergerakan nilai PDRB untuk setiap provinsinya. Didapatkan bahwa pola pergerakan nilai PDRB pada setiap provinsi hampir identik dimana terjadi peningkatan signifikan sampai tahun 2019 lalu mengalami penurunan pada tahun 2020. Lalu mengalami peningkatan lagi sampai tahun 2023. Didapatkan pula bahwa setiap tahunnya, DKI Jakarta menduduki peringkat pertama.

B. *Data Preprocessing*

Data preprocessing adalah suatu tahap persiapan data yang dilakukan sebelum proses analisis untuk mengubah data mentah menjadi data yang siap olah. Data mentah (*raw data*) mungkin masih mengandung ketidakkonsistenan, redudansi, kesalahan *input*, atapun *missing value*. Tahapan ini memastikan bahwa data yang akan dianalisis sudah bersih dari kesalahan dan sudah seragam. *Data preprocessing* juga termasuk melakukan transformasi data untuk meningkatkan akurasi model yang akan dibuat nantinya (Pedregosa, 2011).

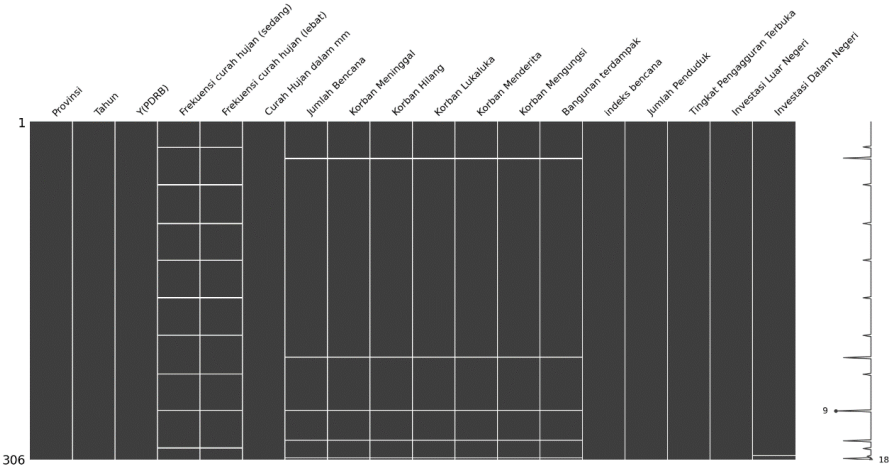
Data preprocessing yang dilakukan dalam penelitian ini menggunakan metode `Pipeline` yang akan meminimalkan kemungkinan *data leakage* atau kebocoran informasi data *test* pada data *train*. Selain itu, `Pipeline` digunakan untuk

meningkatkan efektivitas karena memproses data dalam bentuk *batch* dan mengurangi redundansi sintaks. Sebelum memasuki `Pipeline`, data yang dimiliki dilakukan pemisahan antara data *train* dan data *test*. Diambil seluruh observasi yang terjadi pada tahun 2023 sebagai data *test* sedangkan observasi pada tahun 2015-2022 sebagai data *train*. Pemilihan cara *splitting* data seperti ini dikarenakan pembagian bedasarkan waktu akan memberikan gambaran yang lebih realistis dan general tentang kemampuan model dalam melakukan prediksi di masa depan.

Beberapa metode *preprocessing* data dimasukkan ke dalam `Pipeline` sesuai dengan kondisi data pada penelitian ini. Beberapa di antara metode tersebut dirincikan sebagai berikut.

1. Imputasi *Missing Value*

   Pertama-tama dilakukan identifikasi missing value pada data penelitian sebelum splitting. Ditemukan bahwa beberapa kolom memiliki missing value yang tidak terjadi secara acak. Hal ini terlihat pada Gambar 9 yang menunjukkan bahwa pada tahun dan/atau provinsi tertentu, terdapat beberapa kolom yang secara bersamaan tidak memilki nilai.



Gambar 9. *Identifikasi pola missing value pada keseluruhan data*

Missing value pada data ini akan dilakukan imputasi menggunakan `IterativeImputer`. Teknik imputasi ini mengisi *missing values* dengan cara iteratif menggunakan model *machine learning* untuk mengisi nilai yang hilang

dengan menggunakan informasi dari variabel lainnya. Pada penelitian ini, model *machine learning* yang digunakan adalah `RandomForestRegressor`.

Tahapan imputasi ini dibagi menjadi beberapa bagian sesuai dengan pengelompokkan variabel prediktor. Artinya, variabel bencana dikelompokkan menjadi satu untuk mengisi nilai yang hilang pada kolom-kolom variabel bencana, begitu pula dengan kelompok variabel lainnya. Hal ini dilakukan untuk memastikan bahwa nilai yang hilang setiap kolom tersebut diprediksi menggunakan variabel-variabel yang relevan saja.

2. *Data Scaling*

Data Scaling adalah proses mengubah *range* data agar lebih kecil. Tahapan ini diperlukan karena variabel prediktor yang numerik memiliki skala yang berbeda-beda. Hal ini ditunjukkan pada Gambar 10 dimana beberapa variabel mencatat dalam satuan frekuensi dan yang tidak sehingga memiliki skala atau *range* nilai yang sangat berbeda satu sama lainnya. Perbedaan skala variabel prediktor numerik yang jauh berbeda dapat menimbulkan dominasi fitur dimana fitur/variabel yang memiliki rentang nilai besar akan mendominasi pada model apabila tidak ditangani. Akibatnya, model kurang peka dengan variabel yang memiliki rentang nilai lebih kecil.

Variabel	count	min	median	max
Frekuensi curah hujan (sedang)	297.0	6.03	1.786000e+01	51.99
Frekuensi curah hujan (lebat)	297.0	0.00	3.900000e-01	5.37
Curah Hujan dalam mm	306.0	1138,41	2.534300e+03	4883.67
Jumlah Bencana	301.0	1.00	1.900000e+01	256.00
Korban Meninggal	301.0	0.00	1.000000e+00	187.00
Korban Hilang	301.0	0.00	0.000000e+00	88.00
Korban Lukaluka	301.0	0.00	0.000000e+00	4593.00
Korban Menderita	301.0	0,00	2.417900e+04	1440064.00

Korban Mengungsi	301.0	0.00	9.710000e+02	430968.00
Bangunan terdampak	301.0	0.00	8.300000e+01	55706.00
indeks bencana	306.0	60.43	1.452124e+02	190.88
Jumlah Penduduk	306.0	565805.00	4.096690e+06	49899992.00
Tingkat Pengagguran Terbuka	306.0	1.40	4.785000e+00	10.95
Investasi Luar Negeri	306.0	2.00	4.171000e+02	8283.70
Investasi Dalam Negeri	305.0	8.80	5.015500e+03	95202.10

Metode *data scaling* yang digunakan pada penelitian ini adalah `MinMaxScaler`. Metode ini mengubah skala data numerik agar bernilai di antara 0 dan 1. Alasan penggunaan `MinMaxScaler` sebagai metode *scaling* adalah karena variabel numerik tidak berdistribusi secara normal. Dengan kata lain, metode normalisasi lebih cocok digunakan sedangkan standarisasi tidak dapat diterapkan. Metode normalisasi data yang paling populer digunakan adalah *min-max scaling* dikarenakan proses yang sederhana dalam menyeragamkan skala data.

Normalisasi dengan metode *min-max* dapat mengubah data sehingga bernilai di antara dua angka yang diinginkan. Secara umum, formula dari normalisasi *min-max* adalah sebagai berikut.

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}} (new\_max - new\_min) + new\ \_min$$

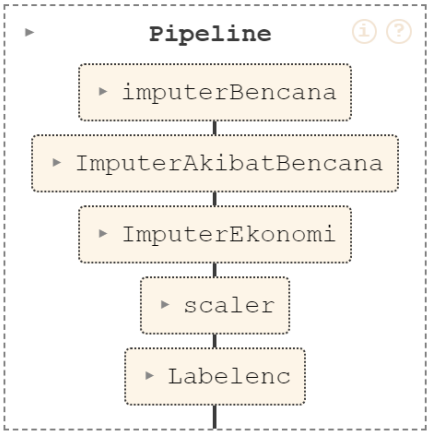
Pada module `sklearn` di Python, fungsi `MinMaxScaler` secara default menetapkan nilai new_max = 1 dan new_min = 0. Artinya, fungsi ini akan mengubah skala data sehingga berada pada rentang nilai 0 sampai 1.

3. *Feature Encoding*

Feature encoding adalah tahapan mengkonversi variabel kategorik menjadi variabel numerik. Tahap ini dilakukan agar data dapat masuk pada algoritma *machine learning* yang sebagian besar membutuhkan *input* dalam format numerik. Penelitian ini melakukan *feature encoding* menggunakan `LabelEncoder`. Variabel

kategorik sebagai prediktor adalah variabel Provinsi. Variabel ini tidak memiliki tingkatan atau strata. Oleh karena itu, metode *feature encoding* dengan `LabelEncoding` lebih tepat digunakan.

Ketiga tahap *data preprocessing* yang telah dijelaskan sebelumnya dikumpulkan secara berurutan pada sebuah `Pipeline`. Urutan tiap tahap ini penting untuk diperhatikan dikarenakan perbedaan urutan akan menghasilkan *output* yang berbeda. Ilustrasi dari alur pada `Pipeline` ditunjukkan pada Gambar 11.



Gambar 10. Ilustrasi Pipeline

C. Hasil Metode Klasik

1.1 Analisis Regresi Data Panel

1.1.1 Uji Hausman

Hal pertama yang akan dilakukan yaitu melakukan Uji Hausman untuk mengetahui apakah terdapat efek random di dalam model panel. Berikut ini adalah uji Hipotesis dari uji Hausman dengan software R.

- Hipotesis

  H_0 : Model merupakan efek random

  H_1 : Model merupakan efek tetap

- Tingkat Signifikansi

  $\alpha = 0{,}05$

- Statistik Uji

chisq	*p-value*
119,38	<2,2e-16

- Daerah Kritik

  H_0 ditolak ketika p-value < α

- Kesimpulan

  Setelah dilakukan diperoleh nilai p – value = < 2.2e-16 < α = 0.05 maka H0 ditolak. Sehingga dapat disimpulkan bahwa model mengandung efek tetap (*fixed effect*).

1.1.2 Uji Breusch Pagan

Uji Breusch-Pagan dilakukan untuk menguji apakah terdapat efek waktu, efek individu, atau efek keduanya pada model regresi panel. Terdapat beberapa aturan pada Uji Breusch-Pagan, yaitu :

- Jika dalam satu model diperoleh kesimpulan "ada efek individu" dan "tidak ada efek waktu", maka terdapat efek individu pada model tersebut.
- Jika dalam satu model diperoleh kesimpulan "tidak ada efek individu" dan "ada efek waktu", maka terdapat efek waktu pada model tersebut.
- Jika dalam satu model diperoleh kesimpulan "ada efek individu" dan "ada efek waktu", maka terdapat efek dua arah pada model tersebut.
- Jika dalam satu model diperoleh kesimpulan "tidak ada efek individu dan waktu", maka model tersebut adalah model regresi pooling

Berikut merupakan pengujian hipotesis dari hasil analisis Uji Breusch-Pagan menggunakana paket program R.

- Hipotesis

  A. Uji 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)

B. Uji 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)

C. Uji 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)

- Tingkat Signifikansi

  $\alpha = 0.05$

- Daerah Kriitk

  H_0 ditolak ketika p-value < α

- Statistik Uji & Kesimpulan

Hipotesis	Statstik Uji	P-Value	Kesimpulan	Kesimpulan Akhir
Individu	Chisq = 158,49	< 2.2e-16	Ada efek individu	Hanya terdapat efek individu
Waktu	Chisq = 0,15885	0.6920	Tidak ada efek waktu	
Individu & Waktu	F=166,52	< 2.2e-16	Ada efek dua arah	

Interpretasi

Untuk mengetahui apakah terdapat apakah terdapat efek waktu, efek individu, atau efek keduanya pada model regresi panel atau tidak, dilakukan uji BreuschPagan dan didapatkan hasil seperti yang tertulis pada tabel diatas. Untuk menguji apakah terdapat efek individu, digunakan hipotesis $H_0 : C = 0, d_t \sim iid, N(0, \sigma_d^2)$ (Tidak terdapat efek individu) vs $H_1 : C \neq 0, d_t \sim iid, N(0, \sigma_d^2)$ (Terdapat efek individu. Kemudian untuk menguji apakah terdapat efek waktu, digunakan hipotesis $H_0 : D = 0, c_i \sim iid, N(0, \sigma_d^2)$ (Tidak terdapat efek waktu) vs $H_1 : D \neq$

$0, c_i \sim iid, N(0, \sigma_d^2)$ (Terdapat efek waktu). Dan terakhir untuk menguji apakah terdapat efek dua arah (efek individu dan waktu), digunakan hipotesis $H_0 : C = 0, D = 0$ (Tidak terdapat efek dua arah pada model) vs $H_1 : C \neq 0 : D \neq 0$ (Terdapat efek dua arah pada model). Pada tingkat signifikansi α = 5%, didapat nilai p-value yang tertulis pada tabel di atas dan daerah penolakan untuk H_0 adalah ditolak jika p-value < α Maka dari Uji Breusch-Pagan tersebut dapat disimpulkan bahwa model efek tetap dengan satu arah yaitu individu.

1.1.3 Uji Wald

Uji Wald dilakukan dengan fungsi *plm()* pada paket program R dengan menspesifikasi model dan jenis efek dari panel. Untuk mengekstrak efek dari setiap kategori digunakan fungsi *fixef()*. Secara umum terdapat dua bentuk pengujian, yaitu pengujian secara simultan dan pengujian variabel secara parsial. Pada uji simultan akan dilihat keseluruhan variabel pada model, apakah model tersebut layak digunakan atau tidak. Sedangkan pada uji parsial, akan diuji masing-masing variabel yang ada pada model. Apabila terdapat variabel yang tidak signifikan, maka variabel tersebut dikeluarkan dari model dan dilakukan pengujian kembali. Berikut merupakan hasil uji inferensi untuk Uji Wald:

- Hipotesis
  a. Uji Simultan/Overall

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

     (model secara simultan tidak layak digunakan)

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

     (model secara simultan signifikan layak digunakan)

  b. Uji Parsial

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

(Koefisien variabel independen tidak signifikan untuk masuk model)

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

(Koefisien variabel independen signifikan untuk masuk model)

- Tingkat Signifikansi

  $\alpha = 0.05$

- Daerah Kritik

  H_0 ditolak ketika nilai p-value < α

- Statistik Uji dan Kesimpulan

Parameter	***P-Value***	***Kesimpulan***
Simultan	<2,22e-16	H_0 ditolak, model secara simultan signifikan layak digunakan
X_1	0,771909	H_0 tidak ditolak, variabel X_1 tidak signifikan pada model
X_2	0,552061	H_0 tidak ditolak, variabel X_2 tidak signifikan pada model
X_3	0,736403	H_0 tidak ditolak, variabel X_3 tidak signifikan pada model
X_4	0,847192	H_0 tidak ditolak, variabel X_4 tidak signifikan pada model
X_5	0,121785	H_0 tidak ditolak, variabel X_5 tidak signifikan pada model
X_6	0,067784	H_0 tidak ditolak, variabel X_6 tidak signifikan pada model
X_7	0,285198	H_0 tidak ditolak, variabel X_7 tidak signifikan pada model
X_8	0,288362	H_0 tidak ditolak, variabel X_8 tidak signifikan pada model
X_9	0,471193	H_0 tidak ditolak, variabel X_9 tidak signifikan pada model
X_{10}	0,070634	H_0 tidak ditolak, variabel X_{10} tidak sigifikan pada model
X_{11}	5,194e-12	H_0 ditolak, variabel X_{11} signifikan pada model

X_{12}	0,001742	H_0 ditolak, variabel X_{12} signifikan pada model
X_{13}	0,453374	H_0 tidak ditolak, variabel X_{13} tidak signifikan pada model
X_{14}	0,213866	H_0 tidak ditolak, variabel X_{14} tidak signifikan pada model
X_{15}	2,2e-16	H_0 ditolak, variabel X_{15} signifikan pada model

Berdasarkan hasil yang didapatkan di atas, model di atas merupakan model efek tetap dengan efek satu arah, dimana dari variabel yang digunakan setelah dilakukan pengujian didapatkan hasil bahwa model layak digunakan. Selain itu didapatkan koefisien variabel yang signifikan adalah Indeks bencana (X_{11}), Jumlah Penduduk (X_{12}), dan Realisasi Investasi Penanaman Modal dalam Negeri (X_{15}).

2.3 *Diagnostic Checking*

Terdapat tiga pengujian yang digunakan untuk diagnostic checking, yaitu uji normalitas korelasi serial dan Uji Heteroskedastisitas. Uji korelasi serial ini bertujuan untuk melihat apakah ada korelasi serial antara komponen galat model, yang mana dapat dilakukan dengan Uji Breusch-Godfrey/Wooldridge. Uji korelasi serial terpenuhi jika tidak ada korelasi serial antar komponen galat. Sedangkan Uji Heteroskedastisitas bertujuan untuk mengetahui apakah model estimasi bersifat tahan (robust) terhadap heteroskedastisitas matriks kovariansi atau tidak, yang dilakukan dengan Heteroskedasticity Robust Covariance Estimator. Asumsi ini terpenuhi jika t-test yang diperoleh konsisten (mempunyai kesimpulan yang sama) pada estimasi model.

1.3.1 Normalitas

- Hipotesis

H_0 : Galat berdistribusi normal

H_1 : Galat tidak berdistribusi normal

- Tingkat Signifikansi

$\alpha = 0.05$

- Daerah Kritik

H_0 ditolak jika nilai p – value < α

- Statistik Uji dan Kesimpulan

D	***p-value***
0,11194	0,002192

Untuk mengetahui apakah pada komponen galat berdistribusi normal, dilakukan uji dengan Kolmogorov-Smirnov dan didapatkan hasil seperti yang tertulis diatas. Digunakan hipotesis utama atau H_0 yaitu galat berdistribusi normal dan hipotesis alternatif atau H_1 yaitu galat tidak berdistribusi normal dengan tingkat signifikansi α = 5% dan daerah penolakan untuk H_0 adalah ditolak jika p – value < α atau dapat dikatakan model merupakan model yang memiliki korelasi serial antar komponen galat. Berdasarkan tabel diatas dapat diketahui bahwa model memiliki galat yang tidak berdistrbusi normal.

2.3.2 Uji Korelasi Serial

- Hipotesis

H_0 : Tidak ada korelasi serial pada komponen galat

H_1 : Ada korelasi serial pada komponen galat

- Tingkat Signifikansi

α = 0.05

- Daerah Kritik

H_0 ditolak jika nilai p – value < α

- Statistik Uji dan Kesimpulan

Chisq	***p-value***
60,793	3,256e-10

Untuk mengetahui apakah terdapat korelasi serial pada komponen galat, dilakukan uji hipotesis dan didapatkan hasil seperti yang tertulis pada tabel diatas. Digunakan hipotesis utama atau H_0 yaitu tidak ada korelasi serial antar komponen galat dan hipotesis alternatif atau H_1 yaitu ada korelasi serial antar komponen galat dengan tingkat signifikansi α = 5% dan daerah penolakan untuk

H_0 adalah ditolak jika p – value < α atau dapat dikatakan model merupakan model yang memiliki korelasi serial antar komponen galat. Berdasarkan tabel diatas dapat diketahui bahwa dari keempat model semuanya terdapat korelasi serial antar komponen galat sehingga semua model tidak memenuhi asumsi ini.

2.3.3 Heteroskedastisitas

- Hipotesis

H_0 : model estimasi tidak bersifat tahan (*robust*) terhadap heteroskedastisitas matriks kovariansi

H_1 : model estimasi bersifat tahan (*robust*) terhadap heteroskedastisitas matriks kovariansi

- Tingkat Signifikansi

  α = 0.05

- Daerah Kritik

  H_0 ditolak jika nilai p – value < α

- Statistik Uji dan Kesimpulan

BP	***p-value***
63,822	5,492e-08

Untuk mengetahui apakah model estimasi bersifat tahan (*robust*) terhadap heteroskedastisitas matriks kovariansi, dilakukan uji hiposesis dan didapatkan hasil seperti yang tertulis pada tabel di atas. Digunakan hipotesis utama H_0 yaitu model estimasi tidak bersifat tahan (*robust*) terhadap heteroskedastisitas matriks kovariansi dan hipotesis alternatif H_1 yaitu model estimasi bersifat tahan (*robust*) terhadap heteroskedastisitas matriks kovariansi dengan tingkat signifikansi α = 5% dan daerah penolakkan untuk H_0 adalah H_0 ditolak apabila p-value < α atau asumsi ini terpenuhi jika t-test yang diperoleh konsisten (memiliki kesimpulan yang sama) pada setiap variabel estimasi model. Diperoleh kesimpulan bahwa model ini tidak konsisten dan asumsi *Heteroskedasticity Robust Covariance* tidak terpenuhi.

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.

D. Hasil Pemodelan Menggunakan *Machine Learning*

Pemodelan menggunakan *machine learning* lebih populer digunakan karena tidak terdapat asumsi-asumsi yang perlu dipenuhi oleh data *input*. Selain itu, model *machine learning* memberikan performa prediksi yang sangat baik. Dalam penelitian ini, akan dilakukan pemilihan model *machine learning* berbasis regresi yang paling baik dengan menggunakan *cross validation*. Selanjutnya, *hyperparameter* pada model *machine learning* terpilih akan dilakukan *tuning* sehingga model terbaik dengan performa maksimal terbentuk. Berikut adalah rincian dari setiap tahap.

1. *Cross Validation*

   Dari ketujuh model, RMSE, MAE, EV, dan R^2 dihitung dan ditentukan kriteria bahwa semakin kecil metriks RMSE, MAE menunjukkan semakin baiknya akurasi prediksi dan semakin besar metriks EV dan R^2 menunjukkan kecocokan yang tinggi antara hasil analisis dan nilai aslinya. Hasil skor CV dengan 5-folds dapat dilihat pada tabel di bawah ini.

Model	***RMSE***	***MAE***	***EV***	***R-square***
XGBRegressor	1.926284e+08	7.485484e+07	0.8296352	0.827479
LinearRegression	1.618418e+08	1.102529e+08	0.8715603	0.866124
ElasticNet	3.735082e+08	2.621791e+08	0.2765083	0.260712
BayesianRidge	4.388316e+08	3.000155e+08	1.374456e-14	-0.028255
GradientBoosting Regressor	1.401976e+08	6.757467e+07	0.9095366	0.908169
CatBoostRegressor	1.483510e+08	7.012480e+07	0.8986391	0.896903
LGBMRegressor	1.587687e+08	9.473549e+07	0.8759354	0.871149

Dari hasil di atas, didapatkan bahwa model *Gradient Boosting Regressor* memenuhi kriteria paling baik dalam memprediksi PDRB. Model tersebut dilakukan pengujian pada data test dan didapatkan nilai RMSE sebesar 1.856451e+08, MAE 1.244616e+08, EV 0.881744, dan R^2 0.86942. Selanjutnya akan dilakukan upaya untuk menaikkan performa model dengan hyperparameter tuning pada model *Gradient Boosting Regressor*.

2. *Hyperparameter Tuning*

Hyperparameter adalah konfigurasi eksternal yang terdapat pada model *machine learning*. Berbeda dengan parameter, *hyperparameter* perlu ditentukan terlenih dahulu sebelum melakukan pemodelan dengan *machine learning*. *Hyperparameter tuning* adalah proses iteratif yang dilakukan untuk mencari kombinasi *hyperparameter* yang menghasilkan performa terbaik. Penelitian ini menggunakan metode random search yang akan mencari kombinasi *hyperparameter* terbaik dari kandidat yang ditentukan dengan mengambil sampel secara random. Kandidat *hyperparameter* pada model GradientBoostingRegressor adalah sebagai berikut.

- `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

Menggunakan fungsi RandomizedSearchCV pada module sklearn, ditentukan jumlah iteration search sebanyak 40 dan jumlah fold pada cv sebanyak 4. Dari kandidat *hyperparameter* yang telah didefinisikan didapatkan kombinasi *hyperparameter* terbaik adalah sebagai berikut.

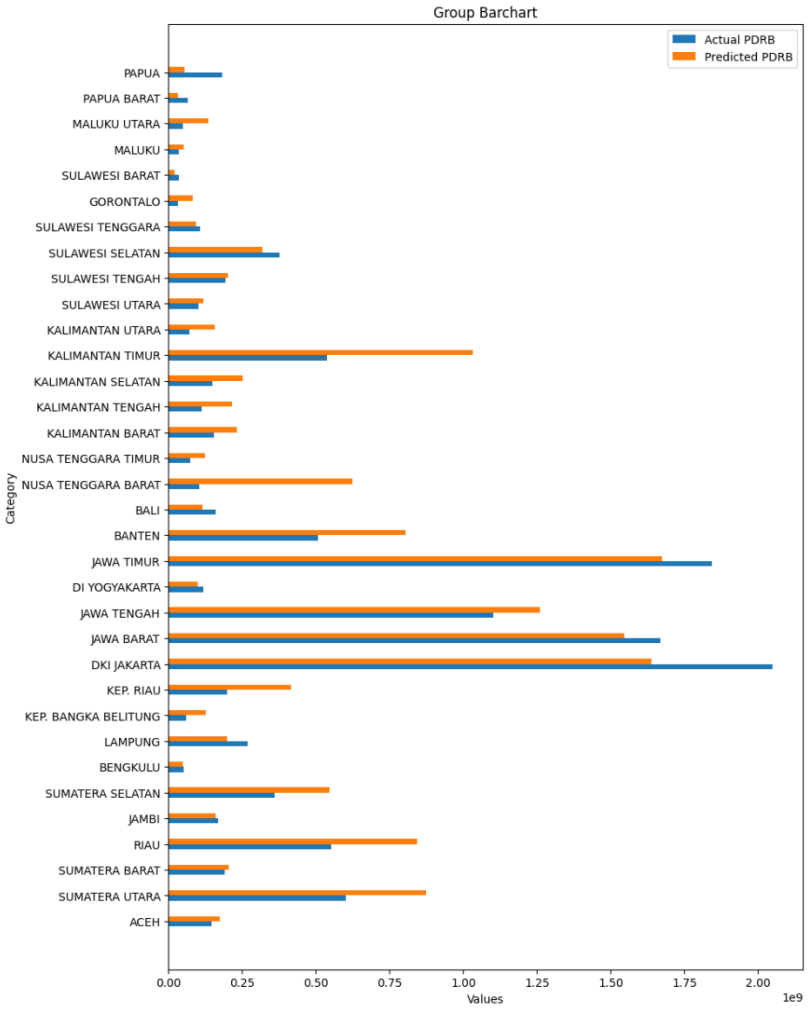
- `n_estimators` = 400
- `learning_rate` = 0.1
- `max_depth` = 3

- `min_samples_split` = 2
- `min_samples_leaf` = 1

Model GradientBoostingRegressor dengan *hyperparameter* terbaik menghasilkan nilai R^2 pada CV sebesar 83,9% sedangkan pada data test menghasilkan nilai R^2 sebesar 86,9%.

E. Prediksi

Model terbaik yang didapatkan yaitu Gradient Boosting Regressor dengan hyperparameter hasil tuning akan digunakan untuk prediksi PDRB pada tahun 2023. Didapatkan hasil prediksi pada gambar di bawah ini.



Gambar 11. Ilustrasi perbandingan hasil prediksi dan data asli

Dari grafik di atas, dapat dilihat perbandingan antara PDRB aktual dan PDRB prediksi untuk berbagai provinsi di Indonesia, dapat diinterpretasikan bahwa terdapat variasi dalam akurasi prediksi model. Beberapa provinsi, seperti Kalimantan Timur dan NTB, menunjukkan prediksi PDRB yang jauh lebih tinggi dibandingkan dengan PDRB aktualnya, sementara DKI Jakarta, Papua, dan Jawa Timur memiliki PDRB aktual yang jauh lebih tinggi dibandingkan dengan prediksi. Provinsi seperti Sulawesi Selatan, Sulawesi Utara, dan Sulawesi Tengah menunjukkan prediksi yang cukup mendekati PDRB aktual, menandakan akurasi model yang lebih baik di daerah-daerah ini. Selain itu, beberapa provinsi dengan nilai PDRB rendah, seperti Bengkulu, DIY, dan Sulawesi Tenggara, menunjukkan konsistensi antara nilai aktual dan prediksi yang rendah. Secara keseluruhan, grafik ini mengindikasikan bahwa model prediksi PDRB memiliki akurasi yang bervariasi di berbagai provinsi, dengan beberapa provinsi menunjukkan perbedaan yang signifikan antara nilai aktual dan prediksi, sehingga memerlukan penyesuaian lebih lanjut dalam model prediksi untuk meningkatkan akurasi, terutama di daerah-daerah dengan perbedaan besar tersebut.

PENUTUP

A. Kesimpulan

Penelitian ini mengusulkan model prediksi yang efektif yang terdiri dari model regresi dengan pendekatan *machine learning* dan model ekonomi klasik, yaitu model regresi linear data panel. Kinerja model *machine learning Gradient Boosting Regressor* dengan *hyperparameter tuning* yang digunakan dalam penelitian ini lebih unggul dibandingkan dengan model lain karena meningkatkan efektivitas prediksi, mengurangi masalah *overfitting*. Model prediksi ini menghasilkan hasil yang sangat baik dalam memprediksi PDRB pada dataset di 34 provinsi di Indonesia dari tahun 2015 hingga 2023, dengan nilai goodness-of-fit sebesar 87%. Selain itu, model ini memiliki kemampuan generalisasi yang tinggi yang dibuktikan pada meningkatnya akurasi yang dihasilkan data train pada data test. Menurut model prediksi, dapat disimpulkan bahwa indikator yang memiliki pengaruh kuat terhadap PDRB untuk 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. Dari penelitian ini, dapat disimpulkan bahwa dari variabel bencana memiliki pegaruh terhadap PDRB Provinsi.

B. Saran

Karena model *machine learning Gradient Bossting Regressor* dengan *hyperparameter tuning* memiliki kemampuan generalisasi yang kuat, maka dari itu sangat direkomendasikan untuk peningkatan kualitas data bencana, terutama indeks risiko bencana dari BNPB untuk meningkatkan akurasi prediksi dengan lebih banyak data. Selain itu melanjutkan dan memperbaiki basis data meteorologi juga sangat penting. Pembuat kebijakan juga dapat lebih memperhatikan daerah dengan indeks bencana banjir dan frekuensi curah hujan yang tinggi karena lebih rentan terhadap bencana banjir.

DAFTAR PUSTAKA

BNPB, B. N. (2022). *Risiko Bencana Indonesia.* Jakarta: BNPB.

Cuello, C. (2023). *Data Integration vs. Data Ingestion*. Diambil kembali dari rivery: https://rivery.io/data-learning-center/data-ingestion-vs-data-integration/

Datamapu. (2024). *Gradient Boost for Regression - Explained*. Diambil kembali dari Datamapu: https://datamapu.com/posts/classical_ml/gradient_boosting_regression/

David, D. (t.thn.). *Web Scraping with Selenium Guide*. Diambil kembali dari brightdata: https://brightdata.com/blog/how-tos/using-selenium-for-web-scraping

Pedregosa, F. (2011). *Scikit-learn: Machine Learning in Python*. Diambil kembali dari scikit-learn: https://scikit-learn.org/stable/index.html

Rosadi, D. (2022). *Pengantar Analisis Data Panel.* Yogyakarta: CV. Meugah Pritindo.

WorldBank. (2021). *Climate Change Knowledge Portal*. Diambil kembali dari World Bank Open Data: https://climateknowledgeportal.worldbank.org/download-data#htab-1497

WRI, W. R. (2022). *World Risk Report 2022.* Jerman: Bündnis Entwicklung Hilft Ruhr University Bochum.

LAMPIRAN

Link Sintaks : Drive Syntax

Link Dataset:

i. inaRISK
ii. DIBI BNPB
iii. Climate Change Knowledge Portal World Bank Open Data
iv. BPS