

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

KAJIAN PERFORMA PENDUGA SELANG PREDIKSI HASIL PEMODELAN *RANDOM FOREST* PADA DATA YANG PEUBAH PENJELASNYA SALING BERKORELASI

MEISYATUL ILMA



**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR**

**BOGOR
2025**

PERNYATAAN MENGENAI TESIS DAN SUMBER INFORMASI SERTA PELIMPAHAN HAK CIPTA

Dengan ini saya menyatakan bahwa tesis dengan judul “Kajian Performa Penduga Selang Prediksi Hasil Pemodelan *Random Forest* pada Data yang Peubah Penjelasnya Saling Berkorelasi” adalah karya saya dengan arahan dari dosen pembimbing dan belum diajukan dalam bentuk apa pun kepada perguruan tinggi mana pun. Sumber informasi yang berasal atau dikutip dari karya yang diterbitkan maupun tidak diterbitkan dari penulis lain telah disebutkan dalam teks dan dicantumkan dalam Daftar Pustaka di bagian akhir tesis ini.

Dengan ini saya melimpahkan hak cipta dari karya tulis saya kepada Institut Pertanian Bogor.

Bogor, Agustus 2025

Meisyatul Ilma
G1501231073

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

RINGKASAN

MEISYATUL ILMA. Kajian Performa Penduga Selang Prediksi Hasil Pemodelan *Random Forest* pada Data yang Peubah Penjelasnya Saling Berkorelasi. Dibimbing oleh BAGUS SARTONO dan HARI WIJAYANTO.

Korelasi tinggi antar peubah penjelas merupakan tantangan dalam pemodelan regresi prediktif karena dapat menimbulkan multikolinearitas yang berdampak pada kestabilan penduga dan interpretasi model. Algoritma *Random Forest* merupakan metode nonparametrik yang mampu memodelkan hubungan nonlinier serta mengakomodasi struktur data berdimensi tinggi tanpa memerlukan spesifikasi eksplisit terhadap bentuk fungsi regresi maupun asumsi tertentu pada sebaran galat. *Random Forest* menghasilkan penduga titik tanpa disertai informasi mengenai ketidakpastian. Keterbatasan ini mengurangi relevansi penggunaan model dalam konteks analisis yang memerlukan estimasi selang atau tingkat keyakinan terhadap hasil prediksi. Pembentukan penduga selang prediksi penting untuk merepresentasikan variabilitas secara kuantitatif, sehingga hasil prediksi tidak hanya akurat tetapi juga informatif secara statistik.

Penelitian ini bertujuan mengkaji kinerja tiga metode pembentukan penduga selang prediksi berbasis *Random Forest*, yaitu *Out-of-Bag Prediction Interval* (OOB-PI), *Quantile Regression Forest* (QRF), dan *Split Conformal Prediction* (SC). Evaluasi dilakukan melalui dua pendekatan, yaitu pertama adalah studi simulasi yang melibatkan variasi tingkat korelasi antar peubah penjelas, bentuk fungsi rata-rata, sebaran galat, dan ukuran contoh. Masing-masing konfigurasi disimulasikan sebanyak 500 replikasi, dengan dua metrik evaluasi utama yaitu *coverage rate* dan lebar selang prediksi. Hasil menunjukkan bahwa OOB-PI memberikan performa paling efisien dengan lebar selang prediksi yang lebih sempit dan cakupan yang mendekati tingkat kepercayaan yang ditentukan. QRF menghasilkan *coverage rate* yang melebihi tingkat kepercayaan, namun disertai dengan selang prediksi yang relatif lebar. SC memberikan jaminan secara teoritis bahwa probabilitas cakupan selang prediksi sesuai dengan tingkat kepercayaan yang ditentukan, tanpa mensyaratkan asumsi tertentu terhadap sebaran data.

Pendekatan kedua merupakan kajian empiris menggunakan data Survei Sosial Ekonomi Nasional (SUSENAS) Provinsi Jawa Barat tahun 2023. Data ini merepresentasikan kondisi aktual yang sering dijumpai dalam praktik, yakni struktur korelasi tinggi antar peubah penjelas. Analisis difokuskan pada enam kabupaten/kota dengan korelasi maksimum antar peubah penjelas di atas 0,74. Peubah respon yang digunakan adalah pengeluaran konsumsi makanan rumah tangga bulanan, sementara 24 peubah penjelas mencakup berbagai dimensi sosial ekonomi rumah tangga. Model *Random Forest* diimplementasikan dengan skema validasi silang *10-fold cross-validation* dan *tuning hyperparameter* untuk setiap wilayah guna mengoptimalkan konfigurasi model.

Analisis data SUSENAS 2023 pada enam wilayah di Provinsi Jawa Barat menunjukkan adanya perbedaan kinerja antarmetode dalam membentuk penduga selang prediksi. Metode OOB-PI menunjukkan performa paling seimbang, dengan *coverage rate* yang konsisten mendekati tingkat kepercayaan dan lebar selang prediksi yang relatif pendek serta stabil di seluruh wilayah. Metode SC menghasilkan *coverage rate* yang sedikit melebihi tingkat kepercayaan, namun



@Himpunan IPB University

IPB University

efisiensinya menurun akibat lebar selang yang seragam dan kurang adaptif terhadap variasi lokal. Metode QRF menunjukkan performa paling tidak efisien, ditandai dengan *coverage rate* yang berada di bawah taraf kepercayaan serta lebar selang prediksi yang paling lebar, sehingga tidak memenuhi kriteria validitas maupun efisiensi pada data dengan korelasi tinggi.

Hasil kajian empiris konsisten dengan hasil studi simulasi yang menunjukkan bahwa OOB-PI merupakan metode paling efisien dalam membentuk penduga selang prediksi, dengan cakupan yang valid dan lebar selang yang relatif sempit. Konsistensi performa ini menjadikan OOB-PI sebagai pendekatan yang paling direkomendasikan dalam pembentukan penduga selang prediksi pada pemodelan *Random Forest*, khususnya pada data dengan korelasi tinggi antar peubah penjelas. Metode ini memberikan estimasi yang presisi sekaligus menyertakan informasi ketidakpastian dalam bentuk selang prediksi yang dapat dievaluasi secara statistik, sehingga relevan untuk analisis sosial ekonomi.

Kata kunci: *Out-of-Bag Prediction Interval*, penduga selang prediksi, *Quantile Regression Forest*, *Random Forest*, *Split Conformal Prediction*

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

SUMMARY

MEISYATUL ILMA. Performance Analysis of Prediction Interval Estimators in Random Forest Modeling on Data with Correlated Explanatory Variables. Supervised by BAGUS SARTONO and HARI WIJAYANTO.

High correlation among explanatory variables poses a challenge in predictive regression modeling, as it can lead to multicollinearity, affecting both the stability of estimators and model interpretation. The Random Forest algorithm is a nonparametric method capable of modeling nonlinear relationships and accommodating high-dimensional data structures without explicit specification of the regression function form or assumptions regarding the error distribution. However, Random Forest produces point estimates without providing information on the associated uncertainty, which limits its applicability in analyses that require interval estimation or confidence assessment of predictions. Constructing prediction intervals is essential to quantitatively represent variability, ensuring that predictions are not only accurate but also statistically informative.

This study aims to evaluate the performance of three Random Forest-based prediction interval estimation methods: Out-of-Bag Prediction Interval (OOB-PI), Quantile Regression Forest (QRF), and Split Conformal Prediction (SC). The evaluation was conducted using two approaches. The first approach employed a simulation study involving variations in correlation levels among explanatory variables, mean function forms, error distributions, and sample sizes. Each configuration was replicated 500 times, and two primary evaluation metrics were used: coverage rate and prediction interval width. The simulation results indicate that OOB-PI yields the most efficient performance, producing narrower prediction intervals with coverage rates close to the nominal confidence level. QRF achieves coverage rates exceeding the nominal level but at the cost of relatively wide intervals. SC theoretically guarantees that the prediction interval coverage probability matches the specified confidence level without requiring distributional assumptions.

The second approach is an empirical study using data from the 2023 National Socioeconomic Survey (SUSENAS) for West Java Province. This dataset reflects real-world conditions frequently encountered in practice, characterized by high correlations among explanatory variables. The analysis focuses on six regencies/cities with maximum pairwise correlations among explanatory variables exceeding 0.74. The response variable is monthly household food expenditure, while the 24 explanatory variables represent various socioeconomic dimensions of households. Random Forest models were implemented using a 10-fold cross-validation scheme with hyperparameter tuning for each region to optimize model configurations.

The empirical analysis of the 2023 SUSENAS data for the six regions in West Java shows performance differences among the prediction interval estimation methods. OOB-PI demonstrates the most balanced performance, with coverage rates consistently close to the nominal confidence level and relatively short, stable prediction intervals across all regions. SC produces slightly higher-than-nominal coverage rates but lower efficiency due to uniform interval widths that are less adaptive to local variations. QRF exhibits the least efficient performance, with



coverage rates below the nominal level and the widest prediction intervals, failing to meet both validity and efficiency criteria in the presence of high correlation among explanatory variables.

The empirical findings are consistent with the simulation results, confirming that OOB-PI is the most efficient method for constructing prediction intervals, offering valid coverage and relatively narrow intervals. This consistency makes OOB-PI the most recommended approach for Random Forest-based prediction interval estimation, particularly for datasets with high correlations among explanatory variables. The method provides precise estimates while incorporating uncertainty information in the form of statistically evaluable prediction intervals, making it highly relevant for socioeconomic analysis.

Keywords: Out-of-Bag Prediction Interval, prediction interval estimation, Quantile Regression Forest, Random Forest, Split Conformal Prediction.



Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Hak Cipta milik IPB, tahun 2025
Hak Cipta dilindungi Undang-Undang

Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan atau menyebutkan sumbernya. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik, atau tinjauan suatu masalah, dan pengutipan tersebut tidak merugikan kepentingan IPB.

Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apa pun tanpa izin IPB.

KAJIAN PERFORMA PENDUGA SELANG PREDIKSI HASIL PEMODELAN *RANDOM FOREST* PADA DATA YANG PEUBAH PENJELASNYA SALING BERKORELASI

MEISYATUL ILMA

Tesis
sebagai salah satu syarat untuk memperoleh gelar
Magister pada
Program Studi Statistika dan Sains Data

**PROGRAM STUDI MAGISTER STATISTIKA DAN SAINS DATA
SEKOLAH SAINS DATA, MATEMATIKA, DAN INFORMATIKA
INSTITUT PERTANIAN BOGOR**

**BOGOR
2025**



@Hak cipta milik IPB University

IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Penguji Luar Komisi Pembimbing pada Ujian Tesis:
Prof. Dr. Ir. I Made Sumertajaya M.Si





Judul Tesis : Kajian Performa Penduga Selang Prediksi Hasil Pemodelan
Random Forest pada Data yang Peubah Penjelasnya Saling
Berkorelasi.

Nama : Meisyatul Ilma
NIM : G1501231073

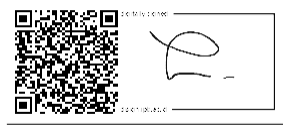
@Hak cipta milik IPB University

Disetujui oleh

Pembimbing 1:
Dr. Bagus Sartono, S.Si., M.Si



Pembimbing 2:
Prof. Dr. Ir. Hari Wijayanto, M.Si



Diketahui oleh

Ketua Program Studi:
Dr. Agus Mohamad Soleh, S.Si., M.T.
NIP. 19750315 199903 1 004



Dekan Sekolah Sains Data, Matematika, dan Informatika
Prof. Dr. Ir. Agus Bueno, M.Si., M.Kom.
NIP. 19660702 199302 1 001



Tanggal Ujian: 6 Agustus 2025

Tanggal Lulus:

PRAKATA

Puji syukur kehadiran Allah *subhanahu wa ta'ala* penulis ucapkan atas limpahan rahmat, karunia, dan hidayah-Nya, sehingga penelitian berjudul “Kajian Performa Penduga Selang Prediksi Hasil Pemodelan *Random Forest* pada Data yang Peubah Penjelasnya Saling Berkorelasi” dapat diselesaikan dengan baik. Penelitian ini disusun sebagai salah satu syarat untuk memperoleh gelar Magister Sains pada Program Studi Statistika dan Sains Data.

Terselesaikannya tesis ini tidak terlepas dari bantuan, bimbingan, dan dukungan dari berbagai pihak. Oleh karena itu, dengan penuh rasa hormat dan tulus, penulis menyampaikan terima kasih kepada:

1. Bapak Dr. Bagus Sartono, S.Si., M.Si. dan Prof. Dr. Ir. Hari Wijayanto, M.Si., selaku komisi pembimbing yang dengan penuh kesabaran dan dedikasi telah meluangkan waktu untuk memberikan bimbingan, arahan, masukan, dan motivasi selama proses penelitian hingga penyusunan tesis ini.
2. Bapak Prof. Dr. Ir. I Made Sumertajaya, M.Si., selaku penguji luar komisi yang telah memberikan saran, kritik, dan masukan yang sangat berharga demi penyempurnaan tesis ini.
3. Badan Pusat Statistik (BPS) yang telah memberikan akses data penelitian sehingga penelitian ini dapat terlaksana.
4. Kedua orang tua, kakak, dan adik, atas doa, kasih sayang, dan dukungan moral yang tidak pernah terputus.
5. Rekan-rekan mahasiswa Program Magister Statistika dan Sains Data angkatan 2023 atas kebersamaan, kerja sama, dan semangat yang telah diberikan selama masa studi.

Penulis menyadari bahwa tesis ini masih memiliki keterbatasan. Namun demikian, besar harapan penulis agar hasil penelitian ini dapat memberikan manfaat bagi pengembangan ilmu pengetahuan dan menjadi rujukan bagi pihak yang membutuhkan.

Bogor, Agustus 2025

Meisyatul Ilma



DAFTAR ISI

DAFTAR TABEL	ix
DAFTAR GAMBAR	ix
DAFTAR LAMPIRAN	x
PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	3
1.3 Tujuan	3
1.4 Manfaat	3
TINJAUAN PUSTAKA	4
2.1 Pemodelan Regresi <i>Random Forest</i>	4
2.2 Penduga Selang Prediksi	5
2.3 <i>Out-of-Bag Prediction Interval</i>	6
2.4 Sifat Asimtotik Penduga Selang OOB-PI	8
2.5 <i>Quantile Regression Forest</i>	9
2.6 <i>Split Conformal Prediction</i>	10
2.7 <i>Cross-Validation</i>	12
III METODOLOGI PENELITIAN	13
3.1 Kerangka Berpikir	13
3.2 Kajian Simulasi	13
3.3 Kajian Data Empiris	17
IV HASIL DAN PEMBAHASAN	21
4.1 Karakteristik Data Simulasi	21
4.2 Karakteristik Data Empiris	23
4.3 Kinerja Metode dalam Kajian Simulasi	26
4.4 Temuan Simulasi	34
4.5 Kinerja Metode dalam Kajian Empiris	35
4.6 Evaluasi Prediksi Selang Ketiga Metode Pada Data Empiris	39
V SIMPULAN DAN SARAN	41
5.1 Simpulan	41
5.2 Saran	41
DAFTAR PUSTAKA	42
LAMPIRAN	44
RIWAYAT HIDUP	55

DAFTAR TABEL

1	Skenario dasar simulasi	15
2	Kabupaten/kota yang dianalisis	17
3	Deskripsi peubah penelitian	18
4	Statistik deskriptif pengeluaran bulanan rumah tangga untuk konsumsi makanan dalam satuan rupiah	24
5	Parameter optimal <i>Random Forest</i> dan rata-rata MAE pada masing-masing Kabupaten/Kota	35

DAFTAR GAMBAR

1	Skema kerja algoritma <i>Random Forest</i> dalam regresi.	5
2	Skema kerja algoritma pembentukan penduga selang prediksi menggunakan metode OOB-PI pada pemodelan <i>Random Forest</i>	7
3	Skema kerja algoritma pembentukan penduga selang prediksi menggunakan metode QRF pada pemodelan <i>Random Forest</i>	10
4	Skema kerja algoritma pembentukan penduga selang prediksi menggunakan metode SC pada pemodelan <i>Random Forest</i>	11
5	Diagram alir tahapan penelitian untuk data simulasi	16
6	Diagram alir tahapan penelitian untuk data empiris	20
7	Struktur korelasi antar peubah penjelas (a) korelasi rendah, (b) korelasi sedang, dan (c) korelasi tinggi	21
8	Pola hubungan antara peubah penjelas dan peubah respon berdasarkan fungsi rata-rata (a) linier, (b) nonlinier, dan (c) nonlinier dengan interaksi	22
9	Sebaran peubah respon Y berdasarkan jenis galat (a) homoskedastik, (b) heteroskedastik, dan (c) <i>heavy-tailed</i>	22
10	Korelasi antar peubah penjelas (a) Kota Cirebon dan (b) Kota Banjar	23
11	Sebaran peubah respon respon Kota Cirebon (a) sebelum transformasi logaritma (b) setelah transformasi logaritma	24
12	Korelasi antara peubah penjelas dan peubah respon (a) Kabupaten Bogor dan (b) Kota Banjar	25
13	<i>Coverage rate</i> pada kondisi galat homoskedastik berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	27
14	Lebar selang pada kondisi galat homoskedastik berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	28
15	<i>Coverage rate</i> pada kondisi galat heteroskedastik berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	30
16	Lebar selang pada kondisi galat heteroskedastik berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	31
17	<i>Coverage rate</i> pada kondisi galat <i>heavy-tailed</i> berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	32
18	Lebar selang pada kondisi galat <i>heavy-tailed</i> berdasarkan fungsi rata-rata, tingkat korelasi, dan ukuran contoh	33

19	Sebaran <i>coverage rate</i> dari tiga metode pembentukan penduga selang selang pada masing-masing wilayah.	36
20	Sebaran lebar selang dari tiga metode pembentukan penduga selang selang pada masing-masing wilayah.	38
21	Selang prediksi dan nilai aktual pengeluaran konsumsi makanan rumah tangga di Kota Cirebon dengan tiga metode penduga selang prediksi	39

DAFTAR LAMPIRAN

1	Korelasi antar peubah penjelas pada data empiris	44
2	Sebaran sebaran peubah respon data empiris	44
3	Korelasi peubah penjelas dengan respon	46
4	Rata-rata <i>coverage rate</i> pada kondisi galat homoskedastik	46
5	Rata-rata MAE kondisi galat homoskedastik	47
6	Koefisien konsistensi lebar selang galat homoskedastik	47
7	Rata rata lebar selang galat homoskedastik	48
8	Rata-rata <i>coverage rate</i> pada kondisi galat heteroskedastik	48
9	Rata-rata MAE kondisi galat heteroskedastik	49
10	Rata rata lebar selang galat heteroskedastik	49
11	Koefisien konsistensi lebar selang galat heteroskedastik	50
12	Rata-rata <i>coverage rate</i> pada kondisi galat <i>heavy tailed</i>	50
13	Rata-rata MAE kondisi galat <i>heavy-tailed</i>	51
14	Koefisien konsistensi lebar selang galat <i>heavy-tailed</i>	51
15	Rata rata lebar selang galat <i>heavy-tailed</i>	52
16	Kombinasi parameter optimal dan nilai rata-rata MAE yang diperoleh untuk setiap wilayah.	52
17	Prediksi masing-masing Kabupaten/Kota	53