

Prediksi Korupsi Menggunakan Indikator Tata Kelola Publik (*Public Governance*) Berbasis *Machine Learning*

Intan Azizah^{1*}, Rona Nisa Sofia Amriza², Wiwit Ratnasari³

^{1,2}Program Studi Sistem Informasi, Universitas Telkom, Kampus Purwokerto, Jl. DI Panjaitan No.128,
Purwokerto 53147, Jawa Tengah, Indonesia

³Program Studi Ilmu Perpustakaan dan Informasi, Fakultas Ilmu Pengetahuan Budaya, Universitas Indonesia,
Depok, 16424, Indonesia

¹ intanazizah@student.telkomuniversity.ac.id, ²ronanisa@telkomuniversity.ac.id, ³wiwitratnasari@ui.ac.id

Submit: 19-12-2025, Revisi: 30-12-2025, Diterima: 22-01-2026, Publikasi: 27-01-2026

Abstract— Corruption is a global problem faced by almost all countries and has a significant impact on development. Corruption is known to hamper economic growth by disrupting production efficiency and reducing innovation. Therefore, early detection of potential corruption is crucial. The purpose of this study is to predict corruption levels using public governance indicators through a machine learning approach. The models used include Random Forest Regression, Decision Tree Regression, XGBoost Regression, and Gradient Boosting Regression. The results show that Random Forest Regression has the best predictive performance, as indicated by an R^2 value of 0.9802, higher than the other models. Furthermore, the results of the feature importance analysis indicate that the control of corruption dimension is the most dominant factor in determining the Corruption Perception Index (CPI) value. The predictive model allows for early warning of potential increases in corruption and helps focus improvements in governance by strengthening corruption control to support sustainable development.

Keywords: Corruption prediction, random forest regression, decision tree regression, XGboost regression, gradient boosting regression

I. PENDAHULUAN

Korupsi merupakan permasalahan yang krusial karena memiliki dampak luas terhadap pembangunan ekonomi dan kualitas pemerintahan. Dampak negatif korupsi dapat dirasakan dalam berbagai aspek kehidupan, termasuk bidang budaya, ekonomi, dan politik [1]. Korupsi merujuk pada penyalahgunaan kekuasaan untuk kepentingan pribadi, memiliki dampak yang merugikan bagi

masyarakat [2]. Korupsi diketahui dapat menghambat pertumbuhan ekonomi dengan menghalangi proses produksi yang efisien serta menurunkan tingkat inovasi [3]. Berbagai bukti menunjukkan bahwa korupsi berdampak negatif terhadap pertumbuhan ekonomi, terutama di negara-negara dengan tingkat investasi yang rendah dan kualitas tata kelola pemerintahan yang lemah [4], [5].

Konsep tata kelola pemerintahan merupakan salah satu variabel penting seiring meningkatnya kompleksitas penyelenggaraan pemerintahan di era digital [1]. Tata kelola pemerintahan mencakup pengelolaan sumber daya publik dan perumusan kebijakan yang berdampak pada masyarakat. Tata kelola pemerintahan yang efektif berperan penting dalam meningkatkan kesejahteraan sosial dan mendorong pembangunan berkelanjutan, mendorong partisipasi aktif masyarakat dalam pengambilan keputusan publik, serta memperkuat kelembagaan [6].

Praktik tata kelola pemerintahan yang efektif berperan penting dalam menekan korupsi. Tata kelola yang baik dapat mendorong sektor publik yang lebih transparan dan akuntabel, sementara tata kelola yang lemah dapat merusak kepercayaan publik serta menurunkan kesejahteraan masyarakat. Negara-negara dengan tata kelola yang lebih baik mempunyai label korupsi yang rendah [7].

Mengingat dampak buruk korupsi yang tidak dapat diabaikan, diperlukan kajian yang mendalam untuk

memahami akar penyebabnya, khususnya yang berkaitan dengan tata kelola publik. Dalam konteks ini, *Worldwide Governance Indicators* (WGI) banyak digunakan sebagai kerangka konseptual untuk menilai kualitas institusi publik dan kinerja pemerintahan suatu negara [1], [3]. WGI terdiri dari enam dimensi utama yang merepresentasikan berbagai aspek tata kelola. Dimensi tersebut meliputi partisipasi dan akuntabilitas (*voice and accountability*), stabilitas politik dan ketiadaan kekerasan (*political stability and absence of violence*), efektivitas pemerintahan (*government effectiveness*), kualitas regulasi (*regulatory quality*), supremasi hukum (*rule of law*), dan pengendalian korupsi (*control of corruption*) [8], [9].

Dimensi partisipasi dan akuntabilitas (*Voice and Accountability*) merefleksikan tingkat keterlibatan masyarakat dalam proses politik. Khususnya dalam pemilihan pemerintah, masyarakat dapat menyampaikan pendapat, membentuk organisasi, serta memperoleh informasi melalui media yang bebas dan independen. Aspek ini sering dikaitkan dengan tingkat transparansi dan akuntabilitas pemerintahan. Selanjutnya, stabilitas politik dan ketiadaan kekerasan (*Political Stability and Absence of Violence*) merepresentasikan tingkat kerentanan suatu negara terhadap ketidakstabilan politik dan tindakan kekerasan yang didorong oleh motif politik, termasuk terorisme, yang berpotensi menghambat keberlanjutan pembangunan dan efektivitas kebijakan publik [8], [9].

Dari perspektif kapasitas institusional, efektivitas pemerintahan (*Government Effectiveness*) menilai kualitas layanan publik, profesionalisme aparatur sipil negara, serta independensi birokrasi dari intervensi politik. Dimensi ini mencerminkan kemampuan pemerintah dalam merancang dan mengimplementasikan kebijakan secara efektif. Sementara itu, kualitas regulasi (*Regulatory Quality*) berfokus pada kemampuan pemerintah dalam menyusun dan menerapkan regulasi yang mendukung iklim usaha serta pengembangan sektor swasta [8], [9].

Aspek penegakan hukum direpresentasikan oleh supremasi hukum (*Rule of Law*). Dimensi ini mengevaluasi tingkat kepercayaan masyarakat terhadap sistem hukum serta kepatuhan pada aturan, termasuk tingkat penegakan kontrak yang baik, perlindungan hak

kepemilikan, dan kinerja lembaga penegak hukum. Terakhir, pengendalian korupsi (*Control of Corruption*) menilai tingkat penggunaannya publik untuk kepentingan pribadi dalam skala kecil dan besar, yang secara luas diakui sebagai faktor penting yang memengaruhi kualitas tata kelola dan kinerja pembangunan [8], [9].

Beberapa penelitian sebelumnya telah menganalisis hubungan antara tata kelola publik (*public governance*) serta tingkat korupsi [1], [10]. Namun, kajian-kajian tersebut umumnya menggunakan metode statistik kuantitatif yang berfokus pada pengujian hubungan antarvariabel, sehingga hanya menjelaskan keterkaitan antara tata kelola publik dan korupsi. Pendekatan tersebut masih memiliki keterbatasan karena belum mampu memberikan gambaran mengenai potensi tingkat korupsi di masa depan.

Oleh karena itu, prediksi korupsi menjadi penting karena dapat berfungsi sebagai peringatan dini (*early warning*) terhadap kemungkinan peningkatan korupsi [11]. Berbeda dengan analisis statistik antarvariabel, pendekatan prediktif memungkinkan pembuat kebijakan untuk mengantisipasi risiko korupsi sebelum dampaknya terjadi, sehingga intervensi kebijakan dapat dirancang secara lebih proaktif dan tepat sasaran.

Hingga saat ini, sebagian besar penelitian terdahulu masih berfokus pada analisis hubungan antara tata kelola publik dan korupsi, tanpa melakukan pendekatan prediktif [1], [10]. Penelitian ini mengisi celah tersebut dengan mengembangkan model prediksi korupsi berbasis indikator tata kelola publik menggunakan metode *machine learning*. *Machine learning* yang diunakan dalam penelitian ini adalah *random forest regression*, *decision tree regression*, *XGboost regression*, dan *gradient boosting regression*, yang telah terbukti menunjukkan kinerja yang baik dalam tugas prediksi. Pendekatan prediktif ini diharapkan dapat mendukung pemerintah dan pemangku kepentingan dalam perencanaan kebijakan jangka menengah dan panjang, khususnya dalam upaya perbaikan tata kelola pemerintahan.

II. LITERATURE REVIEW

Tata kelola publik (*Public Governance*) merujuk pada berbagai struktur, sistem, dan institusi yang dibentuk untuk memungkinkan masyarakat mengelola urusan publik secara efektif, mengalokasikan sumber daya, serta mengoordinasikan tindakan mereka secara terarah [1], [12]. Salah satu kerangka tata kelola publik yang paling dikenal dan dikembangkan oleh World Bank adalah *Worldwide Governance Indicators* (WGI). Kerangka WGI mengukur kualitas tata kelola pemerintahan melalui enam dimensi utama yang merepresentasikan berbagai aspek tata kelola, yaitu partisipasi dan akuntabilitas (*voice and accountability*), stabilitas politik dan ketiadaan kekerasan (*political stability and absence of violence*), efektivitas pemerintahan (*government effectiveness*), kualitas regulasi (*regulatory quality*), supremasi hukum (*rule of law*), serta pengendalian korupsi (*control of corruption*) [8], [9].

Korupsi didefinisikan sebagai penyalahgunaan kekuasaan publik untuk kepentingan pribadi [13], [14]. Porta dan Vannucci [15] berpendapat bahwa korupsi tidak hanya merupakan pelanggaran hukum semata, melainkan sebuah fenomena sistemik yang tertanam dalam interaksi sosial dan politik, yang sering kali menjebak individu dalam praktik-praktik ilegal yang telah terlembagakan. Huntington (1968) mengkarakterisasikan korupsi sebagai gejala dari proses transisi politik di masyarakat berkembang; namun, pandangan ini telah banyak dikritik dalam kajian-kajian kontemporer [17].

Prediksi tingkat korupsi telah dilakukan menggunakan berbagai metode *machine learning*. Sebagai contoh, Thanh et al. [17] mengimplementasikan beberapa algoritma, yaitu *Decision Tree*, *Random Forest*, *Gradient Boosted Trees*, dan *Extreme Gradient Boosting* (*XGBoost*), untuk memprediksi tingkat korupsi. Studi lain oleh Thanh dan Huy [18] berfokus pada prediksi penipuan keuangan dengan menerapkan sejumlah algoritma *machine learning*, termasuk *Logistic Regression*, *k-Nearest Neighbors* (KNN), *Naive Bayes*, *Support Vector Machine* (SVM) dengan kernel linear dan sigmoid, *Decision Tree*, *Random Forest*, *Artificial Neural Network* (ANN), serta *XGBoost*. Temuan dari berbagai penelitian terdahulu menunjukkan bahwa pendekatan *machine*

learning memiliki kinerja yang andal dalam memodelkan dan memprediksi fenomena korupsi maupun kejahatan keuangan. Oleh karena itu, penelitian ini mengimplementasikan beberapa model regresi berbasis *machine learning*, yaitu *Random Forest Regression*, *Decision Tree Regression*, *Extreme Gradient Boosting* (*XGBoost*) *Regression*, dan *Gradient Boosting Regression*, untuk memprediksi Indeks Persepsi Korupsi.

III. METODOLOGI PENELITIAN

Penelitian ini dilaksanakan melalui lima tahapan utama, sebagaimana ditunjukkan pada Gambar 1. Tahapan awal penelitian ini diawali dengan proses pengumpulan data (*data collection*), di mana data *Worldwide Governance Indicators* (WGI) dan *Corruption Perceptions Index* (CPI) diperoleh dari sumber resmi, yaitu *World Bank*. Selanjutnya, kedua dataset tersebut diintegrasikan ke dalam satu basis data terpadu guna memudahkan proses pengolahan dan analisis pada tahapan berikutnya.

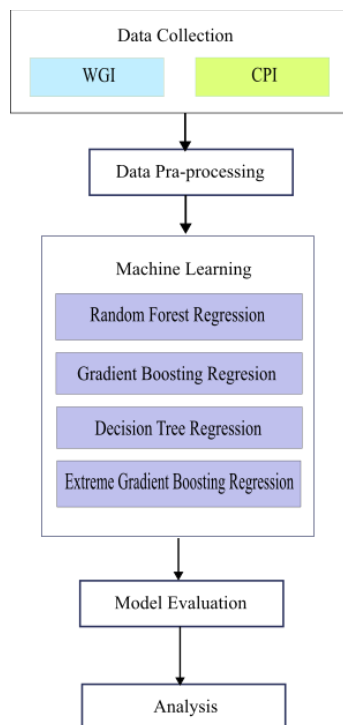
Tahapan kedua dalam penelitian ini adalah data preprocessing. Pada tahap ini dilakukan pembersihan data, yang meliputi identifikasi dan penanganan data duplikat serta nilai hilang (*missing values*) untuk memastikan kualitas dan konsistensi dataset. Selanjutnya, tahapan ketiga adalah pembangunan model prediksi dengan mengimplementasikan empat algoritma *machine learning*. Algoritma yang digunakan meliputi *Random Forest Regression*, *Decision Tree Regression*, *Extreme Gradient Boosting* (*XGBoost*) *Regression*, dan *Gradient Boosting Regression*.

Random Forest Regression memerlukan pemilihan nilai m , yaitu jumlah variabel (sebagai subset dari p variabel prediktor yang tersedia) yang digunakan untuk menentukan keputusan pada setiap simpul (*node*) pohon keputusan [19]. Dalam penelitian ini, penggunaan 300 pohon dipilih untuk mencapai keseimbangan antara performa prediksi dan efisiensi komputasi. Penelitian ini juga menetapkan nilai *random_state* yang tetap, eksperimen dapat diulang dan divalidasi secara konsisten. Selanjutnya seluruh inti prosesor yang tersedia digunakan ($n_jobs = -1$).

Decision Tree Regression merupakan jenis metode *tree-based* yang digunakan untuk memprediksi keluaran

numerik dari variabel dependen [20]. Metode ini memprediksi nilai keluaran menggunakan aturan keputusan berbentuk *if-then*. Secara umum, metode *Decision Tree Regression* melibatkan beberapa tahapan, yaitu menentukan kriteria akurasi prediksi, memilih pemisahan (*splitting*), menetapkan kondisi penghentian pemisahan, serta memilih struktur pohon yang optimal [21].

Extreme Gradient Boosting Regression merupakan salah satu implementasi dari *Gradient Boosting Machine* yang dikenal sebagai salah satu algoritma dengan kinerja terbaik dalam pembelajaran terawasi (*supervised learning*). Algoritma ini digunakan untuk mengatasi permasalahan regresi [22]. Pada pengaturan model, penelitian ini menetapkan nilai *random_state* yang tetap.



Gbr. 1. Metodologi Penelitian

Gradient Boosting Regression merupakan metode *boosting* yang menggabungkan *weak learners* dengan cara secara iteratif memfokuskan pada kesalahan yang dihasilkan pada setiap tahap, hingga diperoleh sebuah *strong learner* yang terbentuk sebagai gabungan dari *weak learners* yang disusun secara berurutan [23].

Tahapan keempat dalam penelitian ini adalah evaluasi model. Kinerja keempat model dievaluasi menggunakan empat metrik pengukuran, yaitu *coefficient of determination* (R^2), *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), dan *Mean Absolute Error* (MAE).

R^2 menunjukkan sejauh mana model mampu menjelaskan variasi pada data aktual. Berikut merupakan formulasi untuk mengukur R^2 :

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}} = \frac{SS_{Reg}}{SS_{Total}} \quad (1)$$

SS_{Res} , SS_{Reg} dan SS_{Total} masing-masing merupakan *sum of squares residual*, *sum of squares regresi*, dan *total sum of squares*.

MSE menilai rata-rata kesalahan kuadrat antara nilai aktual dan hasil prediksi yang dihasilkan oleh model. Berikut merupakan formulasi untuk MSE:

$$MSE = \frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2 \quad (2)$$

merepresentasikan akar kuadrat dari MSE dan memungkinkan interpretasi kesalahan prediksi dalam satuan variabel target. Berikut merupakan formulasi untuk RMSE:

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (X_i - Y_i)^2} \quad (3)$$

Mean Absolute Error (MAE) memiliki nilai absolut rata-rata perbedaan antara nilai aktual dan nilai prediksi yang dihasilkan oleh model. Berikut merupakan formulasi untuk MAE:

$$MAE = \frac{1}{m} \sum_{i=1}^m |X_i - Y_i| \quad (4)$$

m menyatakan jumlah total observasi, merupakan nilai aktual pada observasi ke- i , X_i adalah nilai aktual dan Y_i adalah nilai prediksi yang dihasilkan oleh model. $(X_i - Y_i)$ merupakan selisih antara nilai aktual dan prediksi.

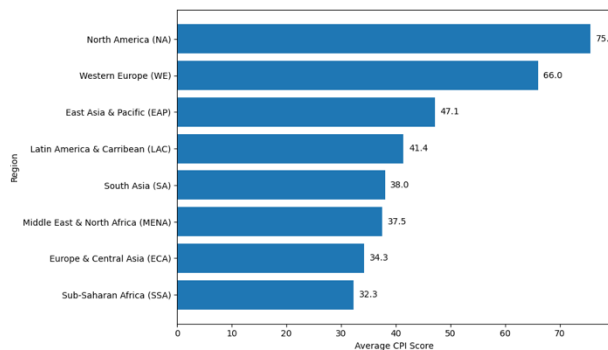
Tahapan kelima dalam penelitian ini adalah analisis hasil prediksi model serta evaluasi *feature importance*. Pada tahap ini dilakukan identifikasi dan analisis terhadap fitur-

fitur yang memiliki pengaruh signifikan terhadap tingkat korupsi berdasarkan indikator tata kelola publik.

IV. HASIL DAN PEMBAHASAN

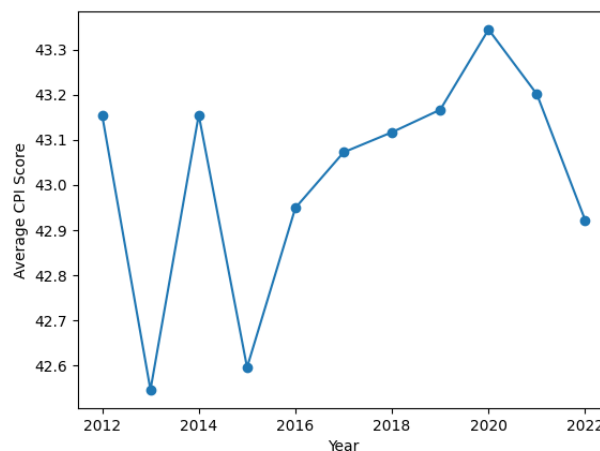
Penelitian ini menggunakan dataset gabungan dari *Worldwide Governance Indicators (WGI)* dan *Corruption Perceptions Index (CPI)* dari tahun 2012 sampai 2022. WGI terdiri dari enam dimensi utama, yaitu partisipasi dan akuntabilitas (*voice and accountability*), stabilitas politik dan ketiadaan kekerasan (*political stability and absence of violence*), efektivitas pemerintahan (*government effectiveness*), kualitas regulasi (*regulatory quality*), supremasi hukum (*rule of law*), dan pengendalian korupsi (*control of corruption*). Dataset berisi data dari 181 negara dan mencakup berbagai wilayah geografis, seperti Asia Timur dan Pasifik, Eropa dan Asia Tengah, Amerika Latin dan Karibia, Timur Tengah dan Afrika Utara, Amerika Utara, Afrika Sub-Sahara, Eropa Barat, serta Asia Selatan.

Indeks Persepsi Korupsi (*Corruption Perceptions Index/CPI*) merupakan alat yang digunakan untuk mengukur tingkat korupsi di berbagai negara di dunia. CPI mengukur tingkat korupsi yang dipersepsikan dalam skala 0 hingga 100, skor yang lebih rendah menunjukkan tingkat korupsi yang lebih tinggi [1], [24]. Gambar 2 menampilkan rata-rata nilai CPI berdasarkan wilayah, yang menunjukkan bahwa tingkat korupsi relatif tertinggi terdapat di wilayah Afrika Sub-Sahara. Selanjutnya, Gambar 3 menggambarkan tren global CPI dari waktu ke waktu. Berdasarkan data tersebut, nilai CPI terendah tercatat pada tahun 2013, sedangkan nilai tertinggi dicapai pada tahun 2020.



Gbr. 1. Rata-rata Indeks Persepsi Korupsi (CPI) berdasarkan Wilayah.

Tahapan awal penelitian ini adalah *pre-processing* data, yang meliputi menghilangkan nilai yang hilang dari data (*missing values*) dan data duplikat. Berdasarkan hasil pemeriksaan, ditemukan sebanyak 33 data dengan nilai hilang pada variabel skor CPI. Selanjutnya, data yang memiliki nilai hilang tersebut dihapus dari dataset sebelum dilakukan analisis lebih lanjut.



Gbr. 2. Perubahan Global Indeks Persepsi Korupsi (CPI) dari Waktu ke Waktu.

Sebelum data dimasukkan ke dalam model, dataset dibagi menjadi data penelitian dan data pengujian. Pada penelitian, 80% data pelatihan dan 20% data pengujian digunakan sisanya merupakan data testing. Variabel independen (*X*) terdiri dari enam indikator tata kelola publik yang diambil dari *Worldwide Governance Indicators (WGI)*, yaitu partisipasi dan akuntabilitas (*voice and accountability*), stabilitas politik dan ketiadaan kekerasan (*political stability and absence of violence*), efektivitas pemerintahan (*government effectiveness*), kualitas regulasi (*regulatory quality*), supremasi hukum (*rule of law*), dan pengendalian korupsi (*control of corruption*). Sementara itu, variabel dependen (*y*) adalah skor Indeks Persepsi Korupsi (CPI) yang merepresentasikan tingkat korupsi suatu negara.

Penelitian ini mengimplemmentasikan empat *machine learning* method yaitu *random forest regression*, *decision tree regression*, *XGboost regression*, dan *gradient boosting regression*. Setiap model memiliki peran penting dalam proses prediksi.

Model *decision tree regression* digunakan untuk menghasilkan aturan yang mudah diinterpretasikan, sehingga membantu memahami proses pengambilan keputusan secara linier. Decision Tree Regression merupakan model non-parametrik dalam *supervised learning* yang digunakan untuk analisis regresi. Model ini didasarkan pada struktur pohon biner yang membagi satu atau lebih simpul (*node*) untuk membentuk sebuah pohon keputusan [25].

Random forest regression merupakan model *ensemble* yang menggabungkan banyak *decision tree* untuk meningkatkan akurasi dan menghindari *overfitting* [24], [26]. Prediksi akhir pada *random forest* diperoleh berdasarkan nilai rata-rata atau pemeringkatan mayoritas dari seluruh pohon. *Random forest regression* sebagai metode *supervised learning* yang berbasis pada pohon regresi, telah terbukti efektif dalam melakukan prediksi pada data yang memiliki interaksi yang kompleks dan bersifat nonlinier [27].

Gradient boosted trees merupakan model *ensemble* lain yang membangun *decision tree* secara berurutan dengan berfokus pada perbaikan kesalahan dari pohon-pohon sebelumnya [28]. Pendekatan ini memungkinkan model menghasilkan prediksi yang semakin kuat pada setiap iterasi dan meningkatkan akurasi model prediksi. XGBoost merupakan algoritma *gradient boosting* yang telah dioptimalkan untuk mencapai kinerja tinggi serta mampu menangani data berskala besar [24].

Tabel 1 menyajikan metrik kinerja dari berbagai model *machine learning*. Model yang dievaluasi meliputi *Decision Tree*, *Random Forest*, *Gradient Boosted Trees*, dan *XGBoost*. Metrik evaluasi yang digunakan adalah R^2 (*coefficient of determination*), RMSE (*Root Mean Squared Error*), MSE (*Mean Squared Error*), dan MAE (*Mean Absolute Error*).

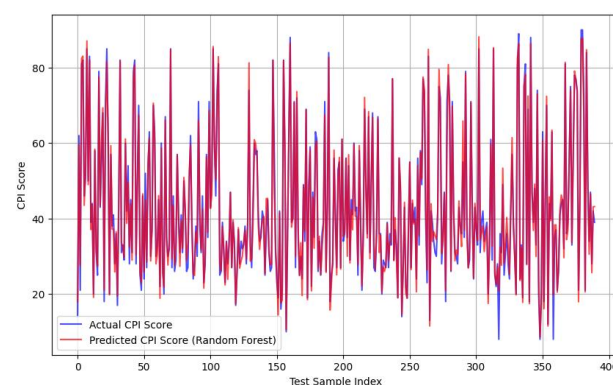
Tabel 1. Evaluasi Performa Model

	R^2	MSE	RMSE	MAE
Random Forest Regression	0.9802	7.1709	2.6778	1.9977
Decision Tree Regression	0.9636	13.1794	3.6303	2.6051
XGBoost Regression	0.9798	7.3064	2.7030	2.6051

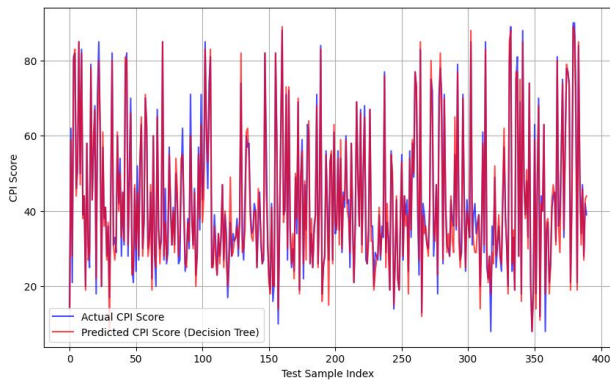
	R^2	MSE	RMSE	MAE
Gradient Boosting Regression	0.9797	7.3464	2.7030	2.6051

Berdasarkan nilai R^2 , seluruh model menunjukkan kemampuan prediksi yang sangat baik dengan nilai di atas 0,96. *Random Forest Regression* memiliki nilai R^2 tertinggi sebesar 0,9802, yang menunjukkan bahwasannya model ini dapat kurang lebih 98,02% variasi data CPI. Nilai R^2 yang tinggi ini mengindikasikan bahwa *Random Forest* memiliki kemampuan terbaik dalam menangkap pola hubungan antara indikator tata kelola publik dan tingkat korupsi. Model *XGBoost* dan *Gradient Boosting* menunjukkan kinerja yang hampir setara, dengan nilai R^2 masing-masing sebesar 0,9798 dan 0,9797, sementara *Decision Tree Regression* memiliki nilai R^2 yang relatif lebih rendah dibandingkan model lainnya.

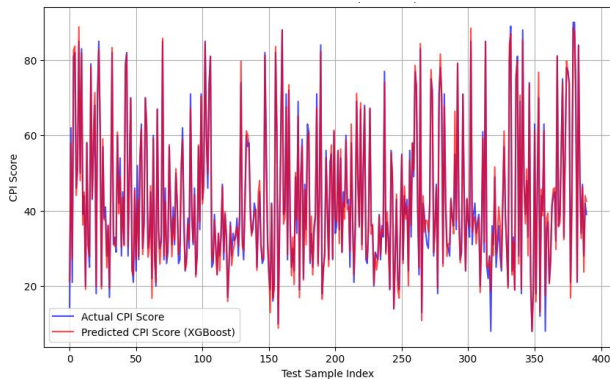
Dari sisi kesalahan prediksi, *Random Forest Regression* juga menunjukkan kinerja terbaik dengan nilai MSE (7,1709), RMSE (2,6778), dan MAE (1,9977) yang paling rendah di antara seluruh model. Nilai kesalahan yang lebih kecil menampilkan bahwa prediksi Random Forest lebih dekat dengan nilai CPI aktual. Sebaliknya, *Decision Tree Regression* memiliki nilai MSE, RMSE, dan MAE yang paling tinggi, yang mengindikasikan bahwa model ini cenderung menghasilkan prediksi yang kurang stabil dan lebih rentan terhadap *overfitting*.



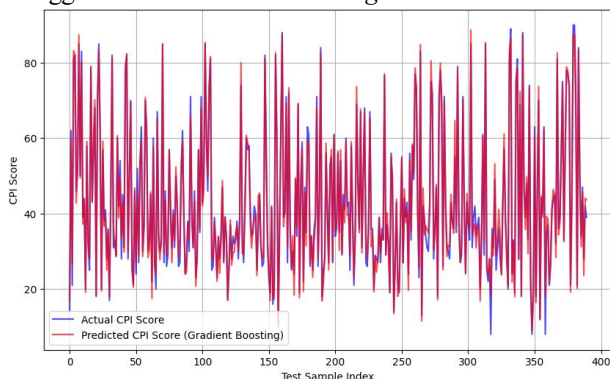
Gbr. 3. Perbandingan Nilai CPI Aktual dan Prediksi menggunakan Model *Random Forest Regression*.



Gbr. 4. Perbandingan Nilai CPI Aktual dan Prediksi menggunakan Model *Decision Tree Regression*.



Gbr. 5. Perbandingan Nilai CPI Aktual dan Prediksi menggunakan Model *XGBoost Regression*

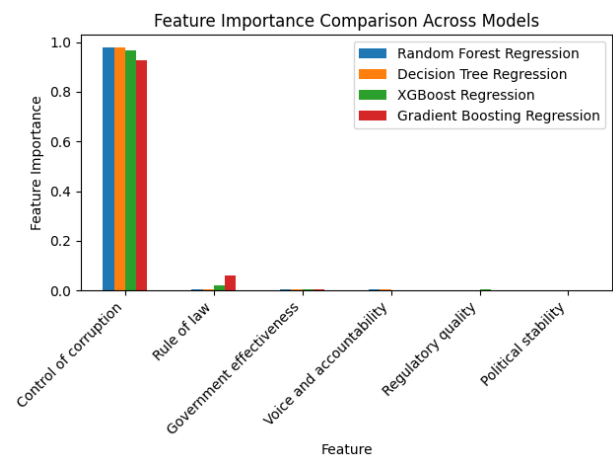


Gbr. 6. Perbandingan Nilai CPI Aktual dan Prediksi menggunakan Model *Gradient Boosting Regression*

Sebagai data pendukung, Gambar 3 hingga Gambar 6 memberikan perbandingan antara nilai Indeks Persepsi Korupsi (CPI) aktual dan nilai CPI hasil prediksi yang dihasilkan oleh empat model machine learning, yaitu *Random Forest Regression*, *Decision Tree Regression*,

XGBoost Regression, dan *Gradient Boosting Regression*. Secara umum, keempat model ini dapat mengikuti pola fluktuasi value CPI aktual dengan cukup baik, yang menunjukkan kemampuan model dalam menangkap hubungan antara indikator tata kelola publik dan tingkat korupsi.

Penelitian ini juga melakukan analisis tingkat kepentingan (*feature importance*) dari masing-masing dimensi *Worldwide Governance Indicators* (WGI) terhadap tingkat persepsi korupsi (CPI). Gambar 7 menunjukkan hasil analisis, dimensi *control of corruption* memiliki tingkat kepentingan tertinggi dibandingkan dengan dimensi WGI lainnya. Temuan ini konsisten pada keempat model *machine learning* yang digunakan, di mana *control of corruption* selalu menunjukkan nilai kepentingan paling tinggi.



Gbr. 7. Perbandingan Tingkat Kepentingan Fitur pada Model.

Tabel 2 menunjukkan detail bobot kepentingan fitur pada berbagai model. Hasil analisis *feature importance* dari enam dimensi *Worldwide Governance Indicators* (WGI) terhadap tingkat persepsi korupsi (CPI) menggunakan empat model machine learning, yaitu *Random Forest Regression*, *Decision Tree Regression*, *XGBoost Regression*, dan *Gradient Boosting Regression*. Nilai *feature importance* menunjukkan kontribusi relatif masing-masing dimensi WGI dalam membentuk prediksi CPI pada setiap model.

Tabel 2. Bobot Kepentingan Fitur pada Berbagai Model

Feature	Importance			
	<i>Random Forest Regression</i>	<i>Decision Tree Regression</i>	<i>XGBoost Regression</i>	<i>Gradient Boosting Regression</i>
Control of corruption	0.9783	0.9799	0.9661	0.9269
Rule of law	0.0069	0.0049	0.0221	0.0623
Government effectiveness	0.0035	0.0034	0.0036	0.0040
Voice and accountability	0.0046	0.0051	0.0021	0.0026
Regulatory quality	0.0033	0.0031	0.0037	0.0024
Political stability	0.0031	0.0033	0.0020	0.0015

Hasil analisis menunjukkan bahwa dimensi *control of corruption* memiliki tingkat kepentingan yang jauh lebih tinggi dibandingkan dimensi lainnya pada keempat model yang digunakan, dengan nilai *feature importance* sebesar 0.9783 pada *Random Forest*, 0.9799 pada *Decision Tree*, 0.9661 pada *XGBoost*, dan 0.9269 pada *Gradient Boosting*. Temuan ini mengindikasikan bahwa dimensi *control of corruption* merupakan faktor yang paling dominan dalam menentukan variasi nilai CPI, serta memiliki pengaruh paling besar terhadap hasil prediksi korupsi.

Sebaliknya, dimensi WGI lainnya, seperti *rule of law*, *government effectiveness*, *voice and accountability*, *regulatory quality*, dan *political stability*, menunjukkan nilai *feature importance* yang relatif rendah pada seluruh model. Meskipun demikian, beberapa variasi kontribusi antar model tetap terlihat, misalnya dimensi *rule of law* yang menunjukkan nilai kepentingan relatif lebih tinggi pada model *Gradient Boosting* dibandingkan model lainnya. Secara keseluruhan, hasil ini menegaskan bahwa meskipun seluruh dimensi WGI berkontribusi dalam pemodelan, pengaruhnya tidak merata, dengan *control of corruption* sebagai faktor utama dalam prediksi tingkat persepsi korupsi.

Dimensi *control of corruption* merefleksikan sejauh mana korupsi, baik itu kecil maupun besar, digunakan untuk kepentingan pribadi dan secara luas diakui sebagai salah satu faktor kunci yang memengaruhi kualitas tata kelola pemerintahan serta kinerja pembangunan. Hasil analisis menunjukkan bahwa *control of corruption* merupakan dimensi yang paling dominan dalam menentukan nilai *indeks persepsi korupsi* (CPI). Temuan ini mengindikasikan bahwa semakin efektif suatu negara

dalam mengendalikan praktik korupsi, semakin rendah tingkat korupsi yang dipersepsikan. Negara yang berhasil menekan berbagai bentuk korupsi, mulai dari suap kecil hingga penggelapan besar, cenderung dipersepsikan lebih bersih dan memiliki nilai CPI yang lebih tinggi.

V. KESIMPULAN

Model ensemble *Random Forest Regression*, memberikan kinerja prediksi yang lebih baik dibandingkan model single learner seperti *Decision Tree*. Temuan ini memperkuat bahwa pendekatan ensemble learning lebih efektif dalam memodelkan hubungan yang kompleks dan nonlinier antara indikator tata kelola publik dan tingkat korupsi.

Hasil penelitian menunjukkan bahwa *control of corruption* merupakan fitur yang paling penting dan paling berpengaruh dalam memprediksi nilai Indeks Persepsi Korupsi (CPI).

Meskipun penelitian ini memberikan kontribusi penting dalam memprediksi tingkat korupsi, beberapa keterbatasan perlu diperhatikan. Pertama, periode data penelitian ini terbatas pada rentang tahun 2012–2022, sehingga belum sepenuhnya mencerminkan kondisi terkini. Akibatnya, penelitian lebih lanjut disarankan untuk menggunakan data yang lebih baru misalnya hingga tahun 2025, agar hasil prediksi menjadi lebih relevan dengan perkembangan terbaru.

Kedua, penelitian ini hanya menggunakan indikator tata kelola publik yang bersumber dari *Worldwide Governance Indicators* (WGI) sebagai variabel independen dalam memprediksi tingkat korupsi dan terbatas pada dimensi partisipasi dan akuntabilitas (*voice and accountability*), stabilitas politik dan ketiadaan kekerasan (*political stability and absence of violence*), efektivitas pemerintahan (*government effectiveness*), kualitas regulasi (*regulatory quality*), supremasi hukum (*rule of law*), dan pengendalian korupsi (*control of corruption*). Penelitian selanjutnya dapat memperluas model dengan menambahkan variabel lain yang berpotensi memengaruhi tingkat korupsi, seperti pendapatan per kapita (*income per capita*), tingkat pendidikan, atau indikator ekonomi makro lainnya, sehingga model prediksi yang dihasilkan menjadi lebih komprehensif dan akurat.

DAFTAR PUSTAKA

- [1] P. S. Koeswayo, S. Handoyo, and D. Abdul Hasyir, "Investigating the Relationship between Public Governance and the Corruption Perception Index," *Cogent Soc. Sci.*, vol. 10, no. 1, p. 2342513, Dec. 2024, doi: 10.1080/23311886.2024.2342513.
- [2] F. Ceschel, A. Hinna, and F. Homborg, "Public Sector Strategies in Curbing Corruption: A Review of the Literature," *Public Organ. Rev.*, vol. 22, no. 3, pp. 571–591, Sep. 2022, doi: 10.1007/s11115-022-00639-4.
- [3] K. Gründler and N. Potrafke, "Corruption and economic growth: New empirical evidence," *Eur. J. Polit. Econ.*, vol. 60, p. 101810, Dec. 2019, doi: 10.1016/j.ejpolco.2019.08.001.
- [4] C.-P. Chang and Y. Hao, "Environmental performance, corruption and economic growth: global evidence using a new data set," *Appl. Econ.*, vol. 49, no. 5, pp. 498–514, Jan. 2017, doi: 10.1080/00036846.2016.1200186.
- [5] A. Cieřlik and Ł. Goczek, "Control of corruption, international investment, and economic growth – Evidence from panel data," *World Dev.*, vol. 103, pp. 323–335, Mar. 2018, doi: 10.1016/j.worlddev.2017.10.028.
- [6] A. Jäntti, H. Paananen, A.-A. Kork, and K. Kurkela, "Towards Interactive Governance: Embedding Citizen Participation in Local Government," *Adm. Soc.*, vol. 55, no. 8, pp. 1529–1554, Sep. 2023, doi: 10.1177/00953997231177220.
- [7] J. Gaventa and R. McGee, "The Impact of Transparency and Accountability Initiatives," *Dev. Policy Rev.*, vol. 31, no. s1, Jul. 2013, doi: 10.1111/dpr.12017.
- [8] S. Handoyo, "Worldwide governance indicators: Cross country data set 2012–2022," *Data Brief*, vol. 51, p. 109814, Dec. 2023, doi: 10.1016/j.dib.2023.109814.
- [9] D. Kaufmann, A. Kraay, and M. Mastruzzi, "The Worldwide Governance Indicators: Methodology and Analytical Issues," *Hague J. Rule Law*, vol. 3, no. 02, pp. 220–246, Sep. 2011, doi: 10.1017/S1876404511200046.
- [10] W. Nor, H. Sofyani, M. Yulastina, and M. Hudaya, "Optimizing intellectual capital for good governance implementation in village government to enhance performance and reduce corruption," *Cogent Bus. Manag.*, vol. 12, no. 1, p. 2574549, Dec. 2025, doi: 10.1080/23311975.2025.2574549.
- [11] G. D. Blasio, A. D'ignazio, and M. Letta, "Predicting Corruption Crimes with Machine Learning. A Study for the Italian Municipalities," 2020, doi: 10.13140/RG.2.2.21886.82245.
- [12] R. A. W. Rhodes, "The New Governance: Governing without Government," *Polit. Stud.*, vol. 44, no. 4, pp. 652–667, Sep. 1996, doi: 10.1111/j.1467-9248.1996.tb01747.x.
- [13] S. Rose-Ackerman, "The Law and Economics of Bribery and Extortion," *Annu. Rev. Law Soc. Sci.*, vol. 6, no. 1, pp. 217–238, Dec. 2010, doi: 10.1146/annurev-lawsocsci-102209-152942.
- [14] A. Paranata, "A Systematic Literature Review of Anti-corruption Policy: A Future Research Agenda in Indonesia," *Public Organ. Rev.*, vol. 25, no. 3, pp. 1181–1214, Sep. 2025, doi: 10.1007/s11115-025-00847-8.
- [15] D. Della Porta and A. Vannucci, "Political Corruption," in *The Wiley-Blackwell Companion to Political Sociology*, 1st ed., E. Amenta, K. Nash, and A. Scott, Eds., Wiley, 2012, pp. 130–143, doi: 10.1002/9781444355093.ch12.
- [16] C. N. Thanh, T. P. Huy, T. P. Hong, and A. B. N. Quoc, "Credit risk prediction with corruption perception index: machine learning approaches," *Cogent Bus. Manag.*, vol. 12, no. 1, p. 2461731, Dec. 2025, doi: 10.1080/23311975.2025.2461731.
- [17] J. Amick, M. Bukovansky, and A. H. Liu, "Presidential Electoral Cycles and Corruption Charges," *J. East Asian Stud.*, vol. 22, no. 2, pp. 281–307, Jul. 2022, doi: 10.1017/jea.2022.9.
- [18] C. N. Thanh and T. P. Huy, "Predicting financial reports fraud by machine learning: the proxy of auditor opinions," *Cogent Bus. Manag.*, vol. 12, no. 1, p. 2510556, Dec. 2025, doi: 10.1080/23311975.2025.2510556.
- [19] P. F. Smith, S. Ganesh, and P. Liu, "A comparison of random forest regression and multiple linear regression for prediction in neuroscience," *J. Neurosci. Methods*, vol. 220, no. 1, pp. 85–91, Oct. 2013, doi: 10.1016/j.jneumeth.2013.08.024.
- [20] S. S. Rathore and S. Kumar, "A Decision Tree Regression based Approach for the Number of Software Faults Prediction," *ACM SIGSOFT Softw. Eng. Notes*, vol. 41, no. 1, pp. 1–6, Feb. 2016, doi: 10.1145/2853073.2853083.
- [21] A. Swetapadma and A. Yadav, "A Novel Decision Tree Regression-Based Fault Distance Estimation Scheme for Transmission Lines," *IEEE Trans. Power Deliv.*, vol. 32, no. 1, pp. 234–245, Feb. 2017, doi: 10.1109/TPWRD.2016.2598553.
- [22] A. Ibrahim Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, Jun. 2021, doi: 10.1016/j.asej.2020.11.011.
- [23] P. Nie, M. Roccotelli, M. P. Fanti, Z. Ming, and Z. Li, "Prediction of home energy consumption based on gradient boosting regression tree," *Energy Rep.*, vol. 7, pp. 1246–1255, Nov. 2021, doi: 10.1016/j.egy.2021.02.006.
- [24] C. Nguyen Thanh, T. P. Huy, T. Pham Hong, and A. Bui Nguyen Quoc, "Credit risk prediction with corruption perception index: machine learning approaches," *Cogent Bus. Manag.*, vol. 12, no. 1, p. 2461731, Dec. 2025, doi: 10.1080/23311975.2025.2461731.
- [25] P. R. Kadavi, C.-W. Lee, and S. Lee, "Landslide-susceptibility mapping in Gangwon-do, South Korea, using logistic regression and decision tree models," *Environ. Earth Sci.*, vol. 78, no. 4, p. 116, Feb. 2019, doi: 10.1007/s12665-019-8119-1.
- [26] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," *Expert Syst. Appl.*, vol. 244, p. 122778, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.
- [27] C. I. Allen Akselrud, "Random forest regression models in ecology: Accounting for messy biological data and producing predictions with uncertainty," *Fish. Res.*, vol. 280, p. 107161, Dec. 2024, doi: 10.1016/j.fishres.2024.107161.
- [28] A. K. George, V. I, and P. Kayal, "Machine learning-enhanced quality minus junk (QMJ) factor and stock returns: Evidence from the indian equity market," *Finance Res. Open*, vol. 1, no. 3, p. 100041, Sep. 2025, doi: 10.1016/j.fmr.2025.100041.