Predicting National Competitiveness Based on Socio-Economic and Environmental Indicators

Introduction

Melalui pendekatan globalisasi, perkembangan dunia tanpa disadari telah semakin bergerak cepat, berbagai negara mendapati sejumlah tantangan yang menghinggapi, salah satu yang terbesar yang mesti ditanggapi ialah perihal persaingan antar negara yang kian meruncing demi mendapatkan suatu ceruk di dunia internasional, tingginya persaingan antar negara menyebabkan dampak pada negara-negara yang memiliki daya saing rendah yaitu ketertinggalan dan menjadi tidak relevan dengan perkembangan zaman. Penyesuaian yang berkelanjutan perlu dilakukan agar suatu negara tidak tertinggal dan mampu beradaptasi dikancah persaingan global.

Saat ini, daya saing nasional merupakan suatu identitas tingkat keberhasilan negara dalam mengelola sumber daya yang dimiliki untuk mempertahankan tingkat produktivitas, pada akhirnya mendorong pembangunan berkelanjutan yang menentukan tingkat kesejahteraan warga negara (Schwab, 2018). Negara berlomba-lomba untuk memperbaiki kondisi pembentuk dari indeks daya saing nasional yang dikeluarkan oleh World Economic Forum. Efek globalisasi tersebut tidak dapat tercapai jika mengabaikan indikator sosial ekonomi dan lingkungan suatu negara (Stiglitz, 2012). Dengan demikian, riset ini menggunakan indikator sosial ekonomi dan lingkungan sebagai variabel independen yang digunakan untuk memprediksi skor daya saing nasional suatu negara. Untuk meminimalkan error pada prediksi skor indeks daya saing nasional digunakan metode analisis regresi dengan model ensemble. Skor indeks daya saing nasional dibutuhkan untuk evaluasi kinerja dari sektor sosial ekonomi dan lingkungan suatu negara untuk mendorong kebijakan yang lebih baik dan lebih cepat.

Related Works

Riset mirip dengan memprediksi daya saing dengan input variabel pengeluaran publik menghasilkan bahwa model artificial neural network merupakan model terbaik dalam memprediksi daya saing (Zaragoza-Ibarra et al., 2021). Metode Analyis Regresi digunakan untuk memprediksi indeks logistic dengan input economic attribute dengan model terbaik yakni artificial neural network (Jomthanachai et al., 2022). Metode yang mirip digunakan pada riset regression analysis of the economic factors of the gross domestic product in the Philippines menggunakan regression analysis dengan linear regression untuk memprediksi gross domestic product dengan nilai error sekitar 7% (Urrutia et al., 2017). Model serupa diterapkan riset dengan judul wisdom of models - business profit prediction using machine learning algorithms untuk memprediksi profit dari sebuah bisnis dengan menggunakan wisdom of models yakni pemilihan model lebih dari satu yakni analysis regresi dengan model linear regression, support vector regression, dan random forest regression (Maheskumar & Revathi, 2021). Dalam pemodelan analysis regresi juga diterapkan pada harga rumah ditandai dengan riset yang berjudul Housing-Price Prediction in Colombia using Machine Learning hal ini meminimalkan error dengan beberapa model yang digunakan antara lain diperoleh adalah 0,25354±0,00699 untuk LightGBM, 0,25296 ±0,00511 untuk Bagging Regressor, dan 0,25312±0,00559 untuk ExtraTree Regressor dengan Bagging Regressor (Ángel Correa Manrique et

al., 2020). Meminimalkan error dengan analysis regresi dapat pula diterapkan untuk memprediksi gaji seseorang disuatu negara, sebagai contoh riset yang berjudul Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations menyatakan tingkat error berkurang signifikan dalam memprediksi gaji ketika menggunakan model artificial neural network dengan penurunan dari 0,38 pada regresi linier menjadi 0,06 dan kesalahan berkurang sekitar 60% (Matbouli & Alghamdi, 2022). Pada riset Evaluation of the Accuracy of Machine Learning Predictions of the Czech Republic's Exports to the China penggunaan Multi Layer Perceptron dianggap model yang efisien digunakan (Suler et al., 2021). Persamaannya ialah menggunakan metode analisis regresi untuk menemukan nilai dari persamaan matematis dari feature inndependen dan perbedaannya sebagian besar menggunakan metode ann sebagai pemilihan model terbaik sementara metode yang dieksplorasi dalam riset ini ialah decisiontree, extratrees, linear regression, randomforest regression, bagging, adaboost, gradientboost, XGboost, dan KNN Regression.

Dataset & Features

Data yang digunakan merupakan data yang dikumpulkan melalui sumber data yakni Bank Dunia, Forum Ekonomi Dunia, Global Competitiveness Report, UNDP, dan macrotrends pada tahun 2019. Data awalnya memiliki 13 Variabel dengan variabel gci merupakan menunjukkan besaran indeks daya saing nasional suatu negara. Adapun variabel dari riset ini ialah, gci (global competitiveness index), negara, gdpp (gross domestic product percapita), fdi (foreign direct investment), inflasi, pengeluaran kesehatan, tingkat pengangguran, emisi karbon, ekspor perkapita, import perkapita, rata-rata lama sekolah, ekspektasi lam sekolah dan kesetaraan gender. Total data keseluruhan sejumlah 141 Negara didunia dan 13 variabel hanya kolom negara saja yang memiliki tipe data kategorik selain itu memiliki tipe data numerik.

Proses pembersihan data ini diawali dengan menghapus variabel yang tidak diperlukan dalam proses pemodelan yakni variabel negara dan ekspektasi lama sekolah, dan selanjutnya data akan melalui proses pemisahan yaitu data training sebesar 80% dari data asli dan data testing sebesar 20% dari data asli. Proses selanjutnya yaitu mengeksplorasi data training untuk dilihat insight dari data tersebut, namun keputusan dari penulis nilai yang cenderung outlier tidak dilakukan perlakuan khusus. Setelah dilakukan eksplorasi data missing value akan diinput dengan nilai median untuk semua data numerik. Terakhir, scalling data dilakukan guna menjaga dataset memiliki rentang nilai tertentu.

Metode

Metode yang akan digunakan pada riset kali ini sebagai eksperimentasi pencarian nilai error terbaik ialah decisiontree, extratrees, linear regression, randomforest regression, bagging, adaboost, gradientboost, XGboost, dan KNN Regression berikut penjelasan singkatnya. Seluruh metode dilakukan untuk meminimalisir mean absolute error (MAE)

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

n: mewakili jumlah pengamatan

yi: mewakili nilai Aktual

ŷi: mewakili nilai Prediksi

dan berdasarkan metode yang digunakan, akan memprediksi nilai output sejumlah satu kolom yang berdasarkan kalkulasi dari nilai input sejumlah 10 kolom

$$f : \mathbb{R}^{10} \rightarrow \mathbb{R}^{1}$$

K nearest neighbors

K nearest neighbors merupakan algoritma sederhana yang berusaha mengestimasi suatu nilai dengan basis titik menggunakan tetangga terdekat. Algoritma KNN menggunakan kemiripan fitur guna memprediksi nilai pada suatu titik data terbaru. Implementasi sederhana dari regresi menggunakan KNN adalah menghitung rata-rata target numerik dari K tetangga terdekat. Ukuran K perlu ditentukan yang dapat dipilih menggunakan cross validation untuk meminimalisir mean absolute errornya.

Decisiontree Regresi

Decisiontree membangun model regresi atau klasifikasi dalam bentuk struktur pohon, yakni dengan memecah suatu data menjadi bagian-bagian atau subset yang lebih kecil, dan saat yang bersamaan decisiontree dikembangkan secara bertahap. Regresi Decisiontree mengamati suatu fitur objek dan melatih model untuk memprediksi output continuous. Hal ini berarti bahwa keluaran/hasil tidak diskrit, yaitu, tidak diwakili hanya oleh kumpulan angka atau nilai yang diketahui dan terpisah. Decisiontree di mana variabel target atau simpul terminal dapat mengambil nilai kontinu (biasanya bilangan real)

Extratrees Regression

Extremely Randomized Trees, atau singkatnya Extratrees, merupakan algoritma machine learning ensemble. Metode ini membuat trees ekstra secara acak dalam sub-sampel kumpulan data untuk meningkatkan prediksi model. Alih-alih menghitung nilai optimal secara lokal menggunakan Gini atau entropi untuk membagi data, algoritme secara acak memilih nilai yang dipisahkan. Hal ini

membuat pohon terdiversifikasi dan tidak berkorelasi. Dengan pendekatan ini, metode ini mengurangi varians. Metode ini merata-ratakan keluaran dari decisiontrees.

Linear Regression

Regresi linier melakukan tugas untuk memprediksi nilai variabel dependen (y) berdasarkan variabel independen (x) yang diberikan). Oleh karena itu, namanya adalah Regresi Linier. Karena regresi linier menunjukkan hubungan yang linier, artinya menemukan bagaimana nilai variabel dependen berubah sesuai dengan nilai variabel independen.

Randomforest Regression

Algoritma Regresi Hutan Acak adalah kelas algoritma Pembelajaran Mesin yang menggunakan kombinasi beberapa pohon keputusan acak yang masing-masing dilatih pada subset data. Penggunaan banyak pohon memberikan stabilitas pada algoritme dan mengurangi varians. Algoritma regresi hutan acak adalah model yang umum digunakan karena kemampuannya bekerja dengan baik untuk data besar dan sebagian besar jenis.

Bagging Regressor

bagging regressor menyesuaikan regressor dasar ke subset acak individu dari dataset asli dan kemudian menggabungkan setiap prediksi (baik dengan majority vote atau rata-rata) untuk mendapatkan prediksi akhir. Dengan menambahkan pengacakan pada proses pembuatan estimator (seperti decisiontree), estimator semacam ini sering dapat digunakan untuk menurunkan varians dari hasil prediksi.

Adaboost Regression

AdaBoost regressor adalah meta-estimator yang dimulai dengan memasang regressor pada dataset asli dan kemudian menyesuaikan salinan regressor tambahan pada dataset yang sama tetapi bobot instance disesuaikan menurut kesalahan prediksi saat ini. Dengan demikian, regressor selanjutnya lebih fokus pada kasus-kasus sulit.

Gradient Boosting

Gradient boost adalah algoritma pembelajaran mesin yang bekerja pada teknik ansambel yang disebut 'Boosting'. Seperti model peningkatan lainnya, Gradient boost secara berurutan menggabungkan banyak pembelajar yang lemah untuk membentuk pembelajar yang kuat. Biasanya Gradient boost menggunakan pohon keputusan sebagai pembelajar yang lemah. Di setiap iterasi, pohon baru yang akan ditambahkan berfokus secara eksplisit pada data yang bertanggung jawab atas kesalahan regresi yang tersisa

XGboost

Dalam konteks regresi, XGBoost adalah jenis algoritma pembelajaran terawasi yang dapat digunakan untuk membuat prediksi pada data numerik kontinu. Ini adalah implementasi dari algoritma

pembelajaran mesin penguat gradien, yang merupakan jenis metode pembelajaran ansambel yang menggabungkan prediksi dari beberapa model yang lebih lemah untuk membuat model yang lebih kuat dan lebih akurat.

XGBoost bekerja dengan membangun ansambel pohon keputusan, di mana setiap pohon dilatih untuk membuat prediksi berdasarkan subkumpulan data yang tersedia. Pohon-pohon tumbuh secara berurutan, dengan setiap pohon belajar dari kesalahan pohon sebelumnya. Prediksi akhir dibuat dengan mengambil rata-rata prediksi dari semua pohon dalam ansambel.

Salah satu keunggulan utama XGBoost adalah kemampuannya menangani data yang hilang dan kumpulan data besar secara efisien. Ini juga memiliki sejumlah hyperparameter yang dapat disetel untuk meningkatkan kinerja model, termasuk kecepatan pembelajaran, kedalaman pohon, dan parameter regularisasi.

Experiment dan Diskusi

Dalam riset ini menggunakan pendekatan analisis regresi, dengan demikian titik titik yang dilalui oleh garis regresi diminimalkan jaraknya dengan garis regresi, maka dari itu pemilihan Mean Absolute Error dirasa cukup cocok untuk dijadikan metric dalam riset ini.

Kali pertama, menggunakan base line model yakni mengkalkulasi nilai rata-rata pada data train yang menghasilkan mae pada data train sebesar 10.43 dan mae pada data test 8.96. Hal ini dirasa kurang cukup untuk memodelkan persamaan linear riset ini. Eksperimen disambung dengan menggunakan model Linear Regression yang menghasilkan mae sebesar 24.48 pada data train dan 20.38 pada data testing hal ini berbeda dengan riset regresi analisis faktor ekonomi untuk memprediksi gross domestic product di Filipina dengan error sebesar 7% (Urrutia et al., 2017). Eksperiment dilanjutkan dengan menggunakan model decisiontree dan akan dilakukan cross validation terlebih dahulu agar model lebih memahami data dan menghindari terjadinya overfitting, cross validation dilakukan dengan 7 fold dan parameter maksimal depth 1 hingga 10 menghasilkan mae pada data train sebesar 1.2 dan mae pada cv sebesar 3.1. Eksperimen dilanjut dengan menggunakan model randomforest, sebelum dilakukan cross validation penulis melakukan modeling dengan menggunakan mode default, yakni menghasilkan mae sebesar mae pada data train sebesar 1.196 dan data testing sebesar 2.28, hal ini dilanjutkan dengan eksperimen dengan cross validation menghasilkan mae data training sebesar data train 1.19 dan mae cv 3.4 proses yang sama dilakukan pada riset optimasi prediksi profit bisnis dengan menggunakan random forest (Maheskumar & Revathi, 2021). Dan Model KNN default menghasilkan mae train sebesar 3.2 dan mae test sebesar 2.8 dan ketika dilakukan cross validation yang memiliki fold sebesar 5 menghasilkan mae train 3.466897dan mae cv sebesar 0.8

Eksperimentasi dilanjut dengan menggunakan Extratrees yang sebelumnya dilakukan cross validation terlebih dahulu dengan jumlah fold 3 yang menghasilkan antara lain mae train sebesar 0.15 dan mae cv sebesar 3.5. Model bagging turut andil dalam proses eksperimentasi dengan nilai mae prediksi default sebesar 1.3 untuk data train dam jika dilakukan cross validation dengan 5 fold menghasilkan mae train sebesar 1.2 dan mae cv sebesar 3.6 proses serupa dilakukan lebih dahulu

oleh riset prediksi harga rumah di Colombia dengan hasil error extratrees dan bagging regressor sebesar 0,25296 ±0,00511 dan 0,25312±0,00559 (Ángel Correa Manrique et al., 2020).

Penerapan boosting turut digunakan dalam proses eksperimentasi, yang pertama menggunakan adaptive boosting (AdaBoost) dengan cross validation 5 fold, pada metode ini menghasilkan mae pada data train sebesar 2.2 dan mae pada cv sebesar 3.99. Dilanjutkan dengan model GradientBoosting dengan default model menghasilkan mae train sebesar 0.457 dan mae test sebesar 2.44 sedangkan ketika melakukan cross validation dengan 5 fold menghasilkan nilai mae train sebesar 0.88 dan mae cv sebesar 3.6 dan terakhir eksperimentasi mengenai boosting yakni menggunakan model xgboost dengan default menghasilkan mae train sebesar 0.0003 dan mae test sebesar 2.38 dan ketika melakukan cross validation dengan 5 fold menghasilkan mae train sebesar 0.0005 dan mae cv sebesar 3.5.

Berikut rangkuman mae train dan mae cv dari masing masing metode yang telah dilakukan eksperimentasi :

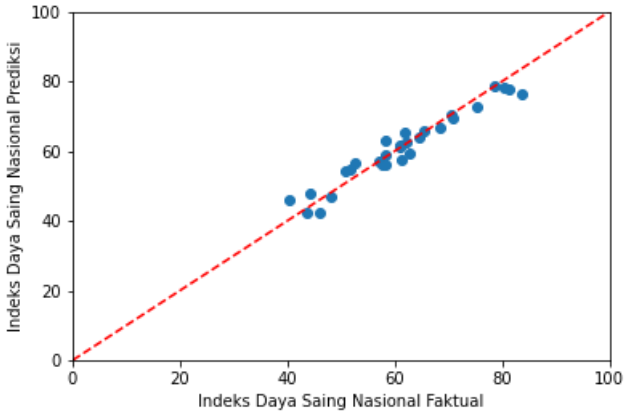
Tabel 1 Komparasi Model

Metode Experimentasi	MAE Train	MAE CV
lin_reg	24.471752	-
decision tree	3.024946	4.574225
ExtraTrees	0.156871	3.466897
bagging	1.285491	3.617310
random forest	1.195786	3.423697
adaBoost	2.187123	3.990832
gradientBoost	0.876317	3.574963
XGboost	0.000504	3.528157
Knn	3.504847	0.82374

Setelah dilakukan eksperimentasi model, sesuai dengan tabel diatas, penulis mendapatkan randomforest sebagai model terpilih untuk memprediksi nilai indeks daya saing nasional dengan nilai mae cv sebesar 3.42 lebih kecil jika dibandingkan dengan hasil mae cv lainnya. Hal ini berbanding terbalik dengan riset yang dilakukan sebelumnya yang meminimalisir mae dengan menggunakan model Neural Network, Multi layer Perceptron, support vector regression, LightGBM, ExtraTree

(Ángel Correa Manrique et al., 2020; Jomthanachai et al., 2022; Maheskumar & Revathi, 2021; Zaragoza-Ibarra et al., 2021)

Gambar 1 Error Analysis



Dari hasil pemodelan regresi dengan menggunakan model randomforest dirasa cukup untuk memprediksi nilai dari global competitiveness index dengan baik. Garis membentang yang diiringi dengan kedekatan titik titik prediksi dan nilai factual dirasa cukup untuk menggambarkan pemodelan dengan menggunakan randomforest.

Conclusion

Dalam riset ini dengan data input sosial ekonomi indikator dan data output global competitiveness indeks mendapatkan hasil randomforest untuk pilihan utama dalam memprediksi nilai output yakni dengan mae sebesar 3.42. Hasil ini dirasa cukup repressentatif untuk memprediksi global competitiveness index yang belakangan ini tidak dilakukan publikasi oleh world economics forum dalam bentuk global competitiveness report. Dengan demikian, setiap negara mampu dengan cepat mengukur daya capaian daya saingnya.

Future Works

Diharapkan menggunakan analisis deep learning dengan menggunakan pendekatan neural network dan untuk variabel inputnya lebih diperbanyak dengan berbagai indikator yang terdapat datanya di world bank atau sumber lain. Dapat pula dilakukan menggunakan analisa Principle Component Analysis yang dilanjutkan dengan analisis regresi, Pendekatan ekonometrika perlu dilakukan untuk mennampilkan pengaruh antar variabel x dan y.

Important Links

https://youtu.be/jc4NzYg8sXc

https://github.com/PrasetyoWidyantoro/Predict-Global-Competitiveness-

Reference

Ángel Correa Manrique, M., Becerra Sierra, O., Otero Gómez, D., Laniado, H., Mateus Carrión, R., & Andres Romero Millan, D. (2020). *Housing-Price Prediction in Colombia using Machine Learning*. *1*.

Jomthanachai, S., Wong, W. P., & Khaw, K. W. (2022). An application of machine learning regression to feature selection: a study of logistics performance and economic attribute. *Neural Computing and Applications*, *34*(18), 15781–15805. https://doi.org/10.1007/s00521-022-07266-6

Maheskumar, V., & Revathi, M. (2021). *WISDOM OF MODELS - BUSINESS PROFIT PREDICTION USING MACHINE LEARNING ALGORITHMS*. *04*, 2391–2396.

Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. *Information (Switzerland)*, *13*(10), 1–14. https://doi.org/10.3390/info13100495

Schwab, K. (2018). The Global Competitiveness Report. In *World Economic Forum*.

Stiglitz, J. E. (2012). *Kegagalan globalisasi dan lembaga-lembaga keuangan Internasional*. Ina Publikatama.

Suler, P., Rowland, Z., & Krulicky, T. (2021). Evaluation of the Accuracy of Machine Learning Predictions of the Czech Republic's Exports to the China. *Journal of Risk and Financial Management*, *14*(2), 76. https://doi.org/10.3390/jrfm14020076

Urrutia, J. D., Tampis, R. L., & Office, M. (2017). *Special Issue REGRESSION ANALYSIS OF THE ECONOMIC FACTORS OF THE GROSS*.

Zaragoza-Ibarra, A., Alfaro-Calderón, G. G., Alfaro-García, V. G., Ornelas-Tellez, F., & Gómez-Monge, R. (2021). A machine learning model of national competitiveness with regional statistics of public expenditure. *Computational and Mathematical Organization Theory*, *27*(4), 451–468. https://doi.org/10.1007/s10588-021-09338-9