ANALISIS PEMODELAN TINGKAT PENGANGGURAN TERBUKA DI INDONESIA: STUDI PERBANDINGAN GEOGRAPHICAL AND TEMPORAL WEIGHTED REGRESSION (GTWR), RANDOM FOREST, DAN XGBOOST DENGAN PERSPEKTIF SPATIO-TEMPORAL

Abstrak

Indonesia adalah negara kepulauan terbesar di dunia. Ini membuat pembangunan dan pemerataan sulit dilakukan. Tujuan SGDs ke-delapan Pekerjaan Layak dan Pertumbuhan Ekonomi. Untuk mencapai tujuan tersebut dilakukan analisis untuk melihat kondisi, memodelkan dan mengetahui faktor yang mempengaruhi TPT. Penelitian ini dilakukan menggunakan pemodelan statistika dan machine learning. Berdasarkan hasil analisis kondisi daerah yang berdekatan cenderung memiliki karakteristik yang serupa. Model terbaik yang didapatkan adalah Xgboost yang memiliki R-squared 94.6% dan RMSE sebesar 0.496 untuk memprediksi TPT tahun 2023. Berdasarkan semua model rata-rata lama sekolah memiliki pengaruh terbesar pada TPT, semakin tinggi rata-rata lama sekolah maka TPT juga akan cenderung meningkat. Eksplorasi data dan model juga menunjukan bahwa daerah yang berdekatan memiliki pengaruh atau karakteristik yang serupa atau adanya efek spasial tidak hanya itu eksplorasi data dan model juga menunjukan efek temporal. Saran dari penelitian ini perlu adanya koordinasi dan kerjasama antar daerah agar dapat menyerap tenaga kerja agar Indonesia dapat mencapai tujuan 8 SDGs.

Kata kunci: GTWR , Random Forest, TPT, Tujuan 8 SDGs, Xgboost

PENDAHULUAN

Latar Belakang

Indonesia sebagai negara kepulauan terbesar di dunia menghadapi tantangan unik dalam mencapai pemerataan pembangunan. Kondisi geografis yang terdiri dari ribuan pulau menyebabkan konektivitas terbatas dan perbedaan sumber daya antar pulau yang menghambat pemerataan dan pembangunan. Salah tujuan pembangunan Indonesia adalah Tujuan ke-8 Pekerjaan layak dan Pertumbuhan ekonomi (Bappenas, 2023). Salah satu faktor yang diukur pada tujuan ke-8 adalah tingkat pengguran terbuka (TPT). Penelitian ini bertujuan untuk mengetahui faktor-faktor yang dapat mempengaruhi TPT di Indonesia. Beberapa variabel dipilih yaitu rata-rata lama sekolah, pengeluaran perkapita dan Umur Harapan Hidup. Pemodelan TPT dilakukan menggunakan *geographically*

weigthed temporally regression, random forest dan *Extreeme Gradient Boosting*. Ketiga metode tersebut akan digunakan dan dibandingkan performanya dari segi ketepatan prediksi, dan seberapa mudah interpretasinya.

Tujuan Penelitian

Berdasarkan latar belakang tujuan dari penelitian ini adalah sebagai berikut:

1. Mendeskripsikan tingkat pengangguran terbuka dan pendidikan di kabupaten dan kota di Indonesia tahun 2019-2023
2. Mendapatkan Model terbaik untuk dapat memprediksi tingkat pengangguran terbuka di kabupaten dan kota di Indonesia tahun 2023
3. Mendapatkan faktor yang mempengaruhi tingkat pengangguran terbuka di kabupaten dan kota di Indonesia

METODOLOGI PENELITIAN

Data

Data yang digunakan pada penelitian ini merupakan data sekunder yang berasal dari panitia, berikut adalah variabel yang digunakan pada penelitian ini.

Tabel 1 Variabel Penelitian

Variabel	Keterangan	Skala	Unit Penelitian
Y	Tingkat Pengangguran Terbuka	Rasio	Kabupaten/Kota
X1	Rata-rata lama Sekolah	Rasio	Kabupaten/Kota
X2	Pengeluaran Perkapita	Rasio	Kabupaten/Kota
X3	Umur Harapan Hidup	Rasio	Kabupaten/Kota

Geographically and Temporally Weighted Regression

Geographically Temporally weighted Regression adalah pengembangan dari model *Geographically weightedRegression* dimana digunakan matriks pembobot spatio temporal sehingga dapat mengukur perubahan parameter seiring Lokasi dan waktu (Debataraja et al., 2021).

$$y_i = \beta_0(u_i, v_i, t_i) + \sum_k \beta_k(u_i, v_i)\, x_{itk} + \varepsilon_i \quad (1)$$

$$\hat{\beta}(u_i, v_i, t_i) = (\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i, t_i)\boldsymbol{X})^{-1}\boldsymbol{X}^T \boldsymbol{W}(u_i, v_i, t_i) y \quad (2)$$

Persamaan nomor 3 menunjukan persamaan *Geographically Temporally Weighted Regression* (GTWR) dan persamaan nomor 4 menunjukan penaksir parameternya. Dimana (u_i, v_i, t_i) merupakan longitude, latitude dan waktu pengamatan data ke-i, $W(u_i, v_i, t_i)$ merupakan matriks pembobot spatio temporal , X merupakan variabel bebas dan yi merupakan variabel dependen. Adapun asumsi error memiliki rata-rata 0, berdistribusi normal dan independen

Random Forest

Random Forest adalah metode machine learning yang berdasarkan tree based model. *Random Forest* adalah sebuah metode bisa yang digunakan untuk klasifikasi, regresi, ataupun tujuan lainnya. *Random Forest* bekerja dengan cara membangun lebih dari satu *Decision Tree* secara random saat training. Hasil yang diberikan oleh Random Forest untuk klasifikasi adalah modus dari klasifikasi decision, sedangkan untuk regresi adalah dengan rata-ratanya (Bentéjac et al., 2019). Metode ini diimplementasikan menggunakan framework tidymodel dari R (Kuhn et al., 2024)

Extreeme Gradient Boosting

Extreeme Gradient Boosting atau XGB adalah pengembangan dari metode tree based method melalui boosting. Boosting sendiri merupakan metode dimana Ketika mendapatkan error dilakukan regresi terhadap error tersebut. Ini dilakukan berulang-ulang kali sampai mendapatkan error terkecil, ini yang dinamakan gradient boosting. Berbeda denga metode boosting biasanya pada XGB terdapat regularisasi L1 dan L2 sehingga bisa mengurangi terjadinya overfitting pada model (Bentéjac et al., 2019). Metode ini diimplementasikan menggunakan framework tidymodel dari R (Kuhn et al., 2024)

Shapley Additive exPlanation (Shap) Value

Memahami alasan di balik prediksi model sama pentingnya dengan akurasi prediksi itu sendiri. Namun, model akurat seringkali kompleks dan sulit dipahami. Metode baru telah dikembangkan untuk membantu interpretasi, tetapi hubungan dan pilihan terbaik di antara metode-metode ini tidak selalu jelas. SHAP dapat digunakan sebagai solusi untuk menyatukan dan memperjelas interpretasi prediksi (Nohara et al., 2022). Berikut ini adalah persamaanya.

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} [f_{S \cup \{i\}}(x - S \cup \{i\}) - f(S)]$$

Dimana:

ϕ_i adalah nilai SHAP untuk variabel j, f adalah model, x adalah input, N adalah himpunan semua fitur, S adalah himpunan bagian dari fitur-fitur yang tidak termasuk fitur j dan f(S) adalah prediksi model untuk instance x dengan hanya menggunakan fitur-fitur yang ada di himpunan S (Nohara et al., 2022).

Langkah Analisis

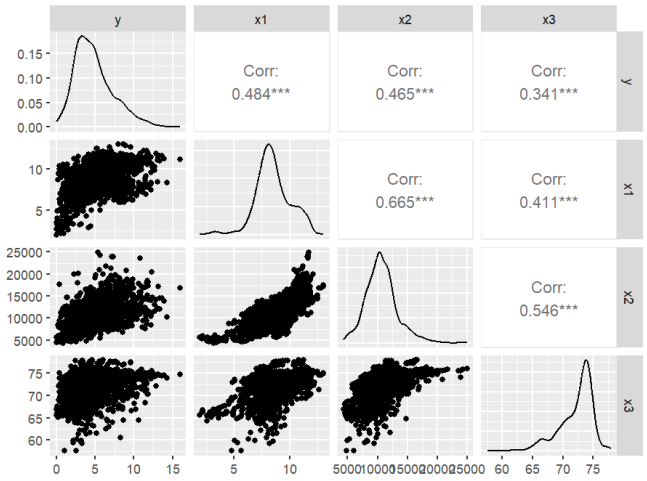
Adapun langkah dari analisis ini pada penelitian ini adalah sebagai berikut ini

Gambar 1 Diagram Langkah Analisis

Berdasarkan gambar 1, langkah pertama dalam penelitian adalah melakukan eksplorasi data. Dilanjutkan dengan pemodelan tanpa efek spasial. Kemudian menguji ada tidaknya efek spasil. Dilanjutkan dengan pemodelan menggunakan efek spasial. Interpretasi faktor yang mempengaruhi tingkat pengangguran terbuka berdasarkan model. Terakhir mengambil kesimpulan dan saran.

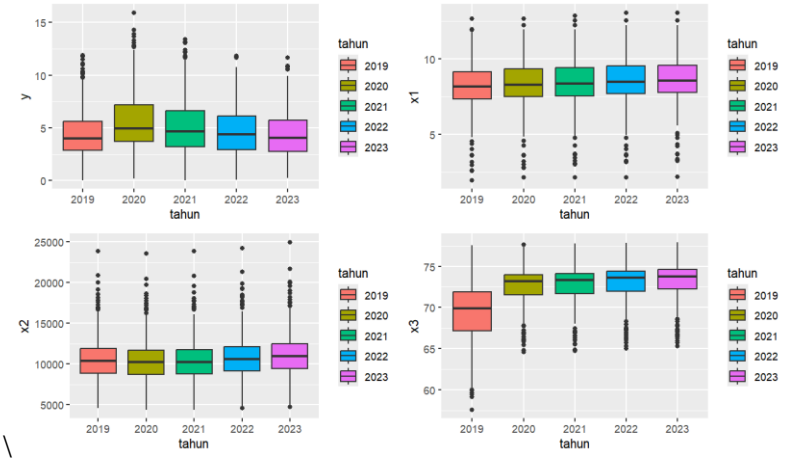
HASIL DAN PEMBAHASAN

Langkah analisis pada penelitian ini dijelaskan pada Gambar 1. Analisis eksplorasi data dan pengujian statistik dilakukan sebelum melakukan pemodelan dengan GTWR, Random Forest, dan *Extreme Gradient Boosting*. Setelah itu, semua model dibandingkan untuk mendapatkan model terbaik. Interpretasi dilakukan untuk model terpilih, dan kemudian dilanjutkan dengan menyusun kesimpulan dan saran.



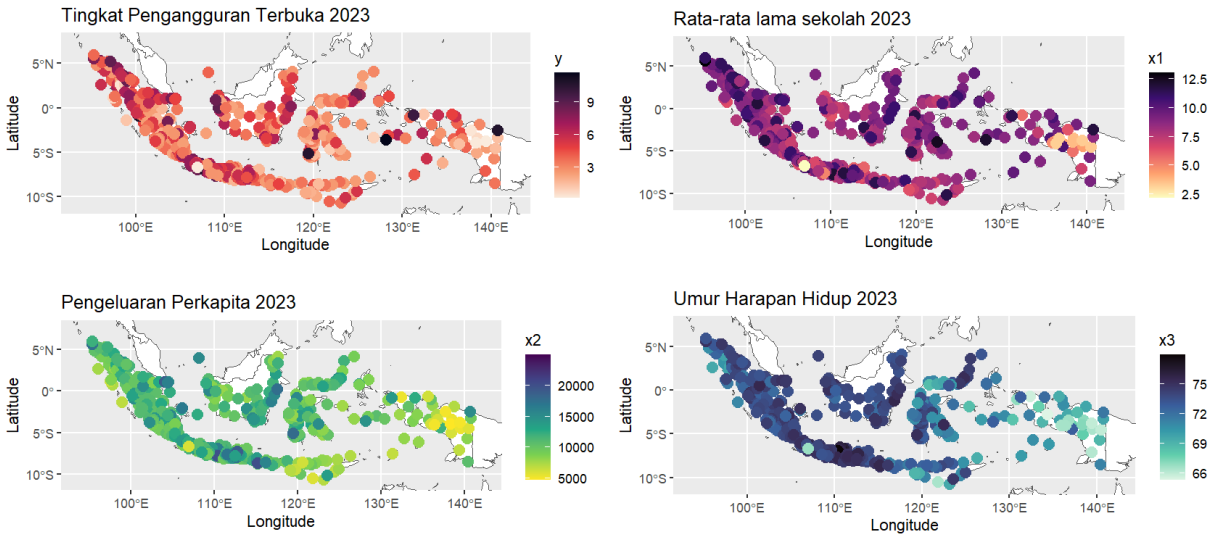
Gambar 2 Matriks Korelasi Variabel

Gambar 2 menunjukkan bahwa tingkat pengangguran terbuka (TPK) memiliki korelasi positif dengan semua variabel lain. Hal ini diduga bahwa jika rata-rata lama sekolah, pengeluaran perkapita, dan angka harapan hidup meningkat, maka TPK juga akan cenderung meningkat.



Gambar 3 Boxplot Variabel Per Tahun

Dapat dilihat berdasarkan Gambar 3 tingkat pengangguran terbuka (y), Rata-rata lama sekolah (x1), pengeluran perkapita (x2) dan umur harapan hidup (x3) mengalami perubahan seiring waktu. Ini dapat menunjukan adanya efek temporal pada data.



Gambar 4 Peta Variabel Tahun 2023

Dapat dilihat berdasarkan gambar 4 tingkat pengangguran terbuka, umur harapan hidup, pengeluaran perkapita dan rata-rata lama sekolah di Indonesia tidak menyebar secara merata. Di pulau jawa, sumatra, Kalimantan dan sulawesi cenderung memiliki tingkat pengguran terbuka, rata-rata lama sekolah, pengeluaran perkapita dan umur harapan hidup yang lebih tinggi dibandingkan Indonesia bagian timur seperti maluku dan papua. Selain itu daerah yang berdekatan juga memiliki kecenderungan nilai variabel yang serupa, ini dapat menjadi indikasi adanya pengaruh spasial pada data. Dilanjutkan dengan pemodelan, berikut adalah rangkuman dari hasil pemodelan.

Setelah melakukan analisis eksplorasi data, berikutnya dilanjutkan dengan pemodelan. Tabel 2 menjelaskan perbandingan kebaikan model dengan R-squared dan RMSE dengan mengabaikan aspek spasial dan temporal.

Tabel 2 Perbandingan model tanpa memperhatikan aspek spatio-temporal

Model	R-Squared	RMSE
Regresi Linear	0.278	2.112
Random Forest	0.408	1.913
Xgboost	0.318	2.062

Dapat dilihat hasil model terbaik adalah dengan menggunakan Random forest yang mampu menjelaskan 40.8% variansi data. Dapat dilihat juga baik regresi linear dan

xgboost hanya mampu menjelaskan 27.8% dan 31.8% dari variansi data. Dilakukan uji breuch pagan untuk melihat pengaruh spasial berikut adalah hasilnya.

H_0 : Tidak ada heteroskedastisitas $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$

H_1 : Ada heteroskedastisitas $\sigma_i^2 \neq \sigma_{\square}^2$ minimal ada satu yang berbeda

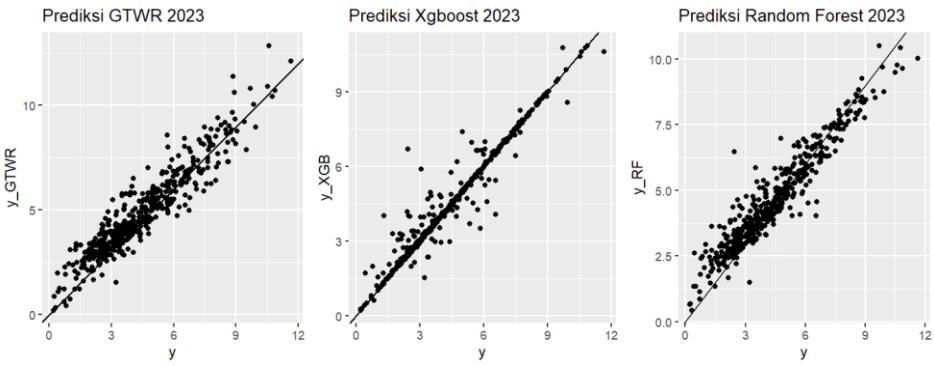
P-value: <2.2e-16

Dapat disimpulkan tolak H0 maka ada heterokedastitas spasial pada model. Dilanjutkan dengan pemodelan menggunakan koordinat untuk mempertimbangkan efek spasial dan temporal, berikut adalah hasilnya:

Tabel 3 Perbandingan model dengan mempertimbangkan aspek spatio-temporal

Model	R-Squared	RMSE
GTWR	0.819	1.508
Random Forest	0.923	0.753
Xgboost	0.962	0.483

Berdasarkan Tabel 3 model terbaik adalah berdasarkan nilai R-Squared dan MSE adalah dengan menggunakan Xgboost yang memiliki nilai R-Squared 96% dan RMSE 0.483, kemudian Random Forest lalu Geographically Temporally *Weighted*Regression . Berikut ini adalah plot perbandingan hasil prediksi dengan data aktual pada tahun 2023.



Gambar 5 Prediksi vs Aktual tahun 2023

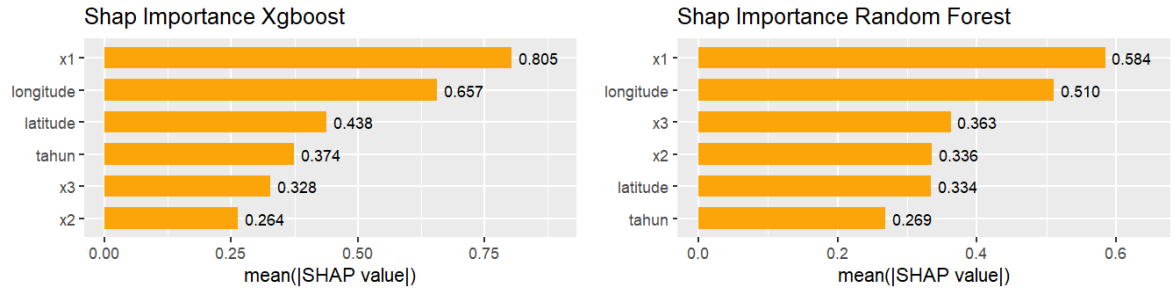
Dapat dilihat berdasarkan hasil prediksi bahwa model sudah dapat mengikuti pola data karena membentuk garis lurus. Berdasarkan ukuran ketepatan didapatkan juga nilai R-squared untuk data tahun 2023 untuk GTWR, *Random Forest*, Xgboost sebesar 88.1%, 91% dan 94.6%, begitu juga nilai RMSE sebesar 0.833, 0.692 dan 0.496. Dapat

disimpulkan model Xgboost adalah model terbaik untuk memprediksi TPT. Berikut ini adalah pengaruh dari variabel pada model.

Tabel 4 Koefisien GTWR

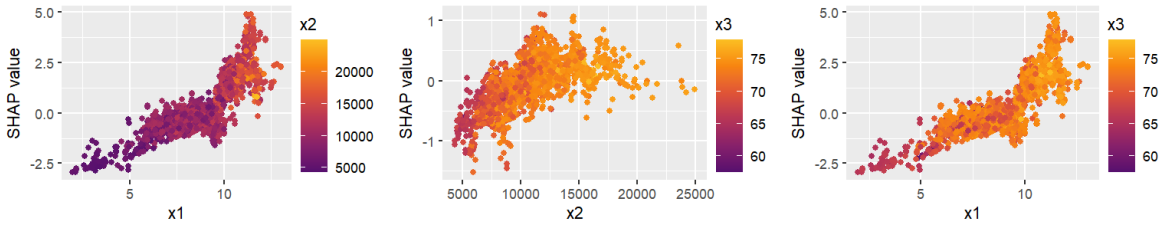
Koefisien	Min	1st Qu.	Median	3rd Qu.	Max.
Intercept	-339.93	-25.085	-10.305	2.475	276.46
Beta1	-2.693	0.00951	0.4458	0.8739	8.556
Beta2	-0.00503	-0.000069	0.00025	0.00051	0.002
Beta3	-4.273	-0.07846	0.1064	0.3292	4.71

Berdasarkan tabel 2 diketahui setidaknya 75% pengamatan mendapatkan pengaruh positif dari variabel x1, untuk variabel x2 dan x3 setidaknya 50% memberikan pengaruh positif terhadap y atau Tingkat pengangguran terbuka. Pengaru positif dari x1 atau rata-rata lama sekolah mungkin terjadi dikarenakan sulitnya menyerap tenaga kerja karena Pendidikan yang meningkat membuat lapangan kerja semakin kompetitif sehingga meningkatkan pengangguran. Berikutnya adalah interpretasi dari model xgboost dan random forest menggunakan SHAP.



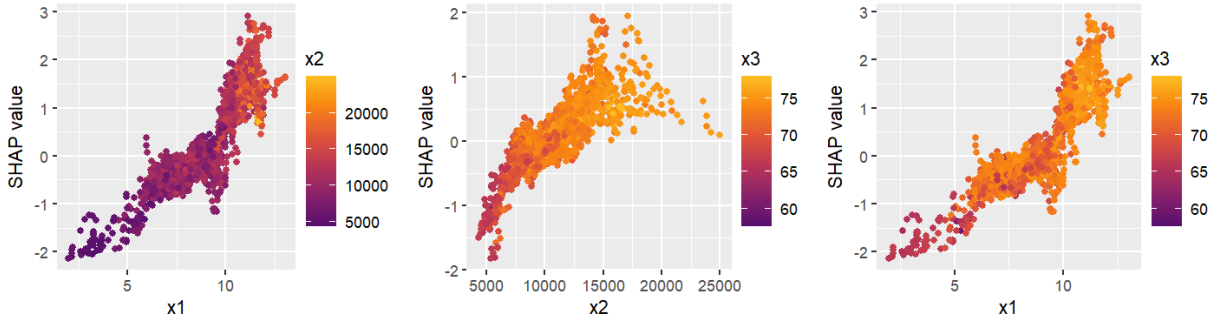
Gambar 6 Feature Importance Model

Berdasarkan gambar 6 dapat dilihat variabel yang memiliki pengaruh terbesar pada model adalah x1. Ini juga sesuai dengan model GTWR di mana 75% pengamatan mendapatkan pengaruh positif dari x1 atau rata-rata lama sekolah. Berdasarkan model xgboost dan random forest juga menunjukan bahwa memang ada pengaruh spasial pada model dengan menggunakan longitude dan latitude. Berikut ini adalah visualisasi untuk melihat pengaruh variabel lebih lanjut dengan menggunakan model Xgboost.



Gambar 7 Shap Dependence Plot Xgboost

Gambar 7 menunjukan pengaruh variabel pada output model. Dapat dilihat ada hubunagan positif pada variabel x1, x2 dan x3 dimana variabel tersebut cenderung naik apabila variabel lain juga naik. Kenaikan variabel x1, x2 maupun x3 juga memiliki pengaruh positif pada nilai prediksi dengan menggunakan model Xgboost.



Gambar 8 Shap Dependence Plot Random Forest

Gambar 8 menunjukan pengaruh variabel pada output model random forest. Dapat dilihat ada hubungan positif pada variabel x1, x2 dan x3 dimana variabel tersebut cenderung naik apabila variabel lain juga naik. Kenaikan variabel x1, x2 maupun x3 juga memiliki pengaruh positif pada nilai prediksi dengan menggunakan model *random forest*.

KESIMPULAN

Berdasarkan hasil pembahasan berikut adalah kesimpulan yang didapat:

1. Tingkap Pengangguran terbuka memiliki hubungan positif dengan rata-rata lama sekolah, pengeluaran perkapita dan angka harapan hidup.
2. Pulau Kalimantan, Sumatra, Jawa, dan Sulawesi memiliki TPT, rata-rata lama sekolah, pengeluaran perkapita dan UHH yang lebih tinggi dibandingkan dengan Indonesia bagian timur.
3. Model terbaik adalah dengan model yang menggunakan koordinat dan tahun pada model, yaitu model Xgboost yang memiliki R-squared 94.6% dan RMSE sebesar 0.496 untuk data tahun 2023 sebesar.

4. Berdasarkan ke-tiga model peningkatan rata-rata lama sekolah memiliki pengaruh terbesar terhadap TPT.

Adapun berikut ini adalah saran dari penelitian ini adalah sebagai berikut:

1. Untuk menurunkan TPT perlu adanya upaya gabungan dari daerah yang berdekatan agar dapat menyerap tenaga kerja.
2. Hubungan positif antara TPT dan Rata-rata lama sekolah dapat menunjukkan bahwa lulusan sekolah belum dapat diserap. Ini mungkin disebabkan kurangnya lapangan kerja atau lulusan sekolah yang belum memiliki standar cukup.

DAFTAR PUSTAKA

Bappenas. (2023). Laporan Pelaksanaan Pencapaian Tujuan Pembangunan Berkelanjutan. In *2023* (Vol. 01).

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2019). *A Comparative Analysis of XGBoost*. 1–20. https://doi.org/10.1007/s10462-020-09896-5

Debataraja, N., Kusnandar, D., Mahalalita, R., & Imro'ah, N. (2021). Penerapan Model Geographically Dan Temporally Weighted Regression Pada Kecelakaan Lalu Lintas. *Jurnal Siger Matematika*, *2*(1), 19–24. https://doi.org/10.23960/jsm.v2i1.2751

Kuhn, M., Wickham, H., & Posit Software, P. (2024). *tidymodels*. 1–6.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2022). *Explanation of Machine Learning Models Using Shapley Additive Explanation and*. *214*(February), 1–7.