Penerapan *Generalized Cross Validation* (GCV) dalam Pemilihan *Bandwidth* Optimal pada Pemodelan Regresi Nonparametrik Kernel (Studi Kasus:Indeks Pembangunan Manusia di Indonesia tahun 2020)

Anggi Adrian Danis[1], Ihsanul Fikri[2], Muhamad Fajri[3], Muhammad Tibri Syofyan[4], Rizqia Salsabila[5]
Program Studi Statistika S1, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Padang
Corresponding author : [1]adriandanis7@gmail.com, [2]ihsanulfikri597@gmail.com, [3]fjrmhd06@gmail.com, [4]tibri.work@gmail.com, [5]rizqiasalsabila2711@gmail.com

ABSTRAK

Pembangunan manusia sangat erat kaitannya dengan pertumbuhan ekonomi. Indeks Pembangunan Manusia (IPM) merupakan salah satu indikator untuk mengukur keberhasilan pembangunan suatu negara. IPM Indonesia secara global tergolong lebih rendah dibandingkan dengan negara-negara di Asia Tenggara lainnya. Oleh karena itu, perlu dilakukan suatu analisis yang tepat untuk memodelkan IPM berdasarkan faktor-faktor yang mempengaruhinya. Analisis di dalam penelitian ini dilakukan dengan menggunakan data IPM tahun 2020 dari 34 provinsi yang merupakan data sekunder yang diperoleh dari Badan Pusat Statistika (BPS), berupa 34 provinsi di Indonesia sebagai unit observasi. Penelitian ini menggunakan fungsi Kernel Gaussian. Fungsi kernel ini digunakan karena fungsi kernel Gaussian lebih mudah dalam perhitungan dan penggunaannya serta lebih sering digunakan dibandingkan dengan fungsi kernel lainnya yang memerlukan syarat dalam pengerjaannya. Metode yang digunakan untuk mendapatkan bandwidth yang optimal adalah dengan meminimumkan nilai Generalized Cross Validation (GCV). GCV merupakan modifikasi dari CV yang didapat dengan meminimumkan fungsi CV.

Kata Kunci : indeks pembangunan manusia, kernel gaussian, regresi nonparametrik kernel

PENDAHULUAN

Indeks Pembangunan Manusia (IPM) atau *Human Development Index* (HDI) adalah pengukuran perbandingan dari harapan hidup, melek huruf, pendidikan dan standar hidup. IPM menjelaskan bagaimana penduduk dapat mengakses hasil pembangunan dalam memperoleh pendapatan, kesehatan, pendidikan, dan sebagainya. IPM dibangun melalui pendekatan tiga dimensi dasar, yaitu kesehatan, pendidikan dan ekonomi.

IPM merupakan hubungan antara manusia dengan pembangunan yang ada disekitarnya, yang mana saling mempengaruhi satu sama lain. Dalam kata lain, terdapat suatu korelasi positif antara nilai IPM dengan derajat keberhasilan pembangunan ekonomi.

IPM diperkenalkan oleh *United Nations Development Programme* (UNDP) pada tahun 1990 dan dipublikasikan secara berkala dalam laporan tahunan *Human Development Report* (HDR). IPM digunakan untuk mengklasifikasikan apakah sebuah negara adalah maju, negara berkembang atau negara terbelakang dan juga untuk mengukur pengaruh dari kebijaksanaan ekonomi terhadap kualitas hidup.

Tujuan utama pembangunan adalah menciptakan lingkungan yang memungkinkan rakyat menikmati umur panjang, sehat dan menjalankan kehidupan yang produktif. Pembangunan manusia menempatkan manusia sebagai tujuan akhir dari pembangunan bukan alat dari pembangunan. Keberhasilan pembangunan manusia dapat dilihat dari seberapa besar permasalahan mendasar masyarakat dapat teratasi.

Estimator kernel adalah pengembangan dari estimator histogram. Estimator ini merupakan estimator linier yang mirip dengan estimator regresi nonparametrik yang lain, perbedaannya hanya karena estimator kernel lebih khusus dalam penggunaan metode *bandwidth*. Hal yang terpenting dalam regresi kernel adalah pemilihan fungsi kernel dan *bandwidth*. Terdapat beberapa fungsi kernel diantaranya kernel Uniform, Triweight, Gaussian, Kuartik, Cosinus dan lainnya.

Kelebihan dari estimator kernel adalah memiliki kemampuan yang baik dalam memodelkan data yang tidak mempunyai pola tertentu, estimator kernel lebih fleksibel, bentuk matematisnya mudah, dan dapat mencapai tingkat kekonvergenan yang relatif cepat. Dari segi komputasi, metode kernel lebih mudah dilakukan dan mudah diimplementasikan.

Estimator kernel *Nadaraya*-Watson merupakan estimasi dengan pendekatan kernel yang bergantung pada dua parameter yaitu *bandwidth* (pemulus) dan fungsi kernel. Pemilihan *bandwidth* optimal untuk mendapatkan kurva regresi yang optimal. Beberapa metode yang dapat digunakan untuk menentukan nilai *bandwidth* optimum yaitu *Cross Validation* (CV), *Akaike Information Criterion* (AIC), *Generalized Cross Validation* (GCV) dan *Bayesian Information Criterion* (BIC).

Penelitian ini menggunakan fungsi Kernel Gaussian. Fungsi kernel ini digunakan karena fungsi kernel Gaussian lebih mudah dalam perhitungan dan penggunaannya serta lebih sering digunakan sedangkan fungs kernel lainnya memerlukan syarat dalam pengerjaannya. Metode yang digunakan untuk mendapatkan bandwidth yang optimal adalah dengan meminimumkan nilai *Generalized Cross Validation* (GCV). GCV merupakan modifikasi dari CV yang didapat dengan meminimumkan fungsi CV.

TINJAUAN PUSTAKA

1. Regresi Nonparametrik

Dalam pendekatan model regresi nonparametrik, kurva regresi tidak diketahui atau tidak terdapat informasi masa lalu yang lengkap tentang bentuk pola datanya. Data diharapkan mencari sendiri bentuk estimasinya sehingga memiliki fleksibilitas yang tinggi. Kurva regresi hanya diasumsikan termuat dalam suatu ruang fungsi yang berdimensi tak hinggadan merupakan fungsi mulus (*smooth*). Estimasi fungsi $z(x_i)$ dilakukan berdasarkan data pengamatan dengan menggunakan teknik *smoothing* tertentu. Ada beberapa teknik *smoothing* yang dapat digurnakan antara lain estimator histogram, kernel, deret orthogonal, spline, k-NN, deret fourier, dan wavelet.

2. Estimator Kernel

Estimator kernel merupakan pengembangan dari estimator histogram. Estimator kernel diperkenalkan oleh Rosenblatt (1956) dan Parzen (1962) sehingga disebut estimator densitas kernel Rosenblatt-Parzen.

Secara umum kernel K dengan *bandwidth* (h) didefinisikan sebagai:

$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$, untuk $-\infty < x < \infty,\ h > 0$ (1)

Serta memenuhi:

(i) $K(x) \geq 0$, untuk semua x

(ii) $\int_{-\infty}^{\infty} K(x)dx = 1$

(iii) $\int_{-\infty}^{\infty} x^2 K(x)dx = \sigma^2 > 0$

(iv) $\int_{-\infty}^{\infty} xK(x)dx = 0$

Maka estimator densitas kernel untuk fungsi densitas $f(x)$ adalah :

$\hat{f}_h = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$ (2)

Dari persamaan (2) terlihat bahwa $\hat{f}_h$ tergantung pada fungsi kernel K dan parameter h. Bentuk bobot kernel ditentukan oleh fungsi kernel K, sedangkan ukuran bobotnya ditentukan oleh parameter pemulus h yang disebut *bandwitdth*. Peran *bandwidth* seperti lebar interval pada histogram.

Beberapa jenis fungsi kernel antara lain :

1. Kernel Uniform : $K(x) = \frac{1}{2} I(|x| \leq 1)$
2. Kernel Triangle : $K(x) = (1 - |x|)\, I(|x| \leq 1)$
3. Kernel Epanechnikov : $K(x) = \frac{3}{4}(1 - x^2)\, I(|x| \leq 1)$
4. Kernel Kuartik : $K(x) = \frac{15}{16}(1 - x^2)^2\, I(|x| \leq 1)$
5. Kernel Triweight : $K(x) = \frac{35}{32}(1 - x^2)^3\, I(|x| \leq 1)$
6. Kernel Cosinus : $K(x) = \frac{\pi}{4}\cos\left(\frac{\pi}{2}x\right) I(|x| \leq 1)$
7. Kernel Gaussian : $K(x) = \frac{1}{\sqrt{2\pi}}\left(\frac{1}{2}(-x^2)\right), -\infty < x < \infty$

dengan I adalah indikator.

3. Regresi Kernel

Regresi kernel adalah teknik statistika nonparametrik untuk mengestimasi fungsi regresi $m(x)$ pada model regresi nonparametrik $y_I = m(x_i) + \varepsilon_i$. Nadaraya dan Watson pada tahun 1964 mendefinisikan estimator regresi kernel sehingga disebut estimator Nadaraya-Watson

$$\hat{m}(x) = \frac{\frac{1}{2}\sum_{i=1}^{n} K_h(x - x_i)y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)}$$

$$\hat{m}(x) = \sum_{i=1}^{n} w_{hi}(x)y_i$$

dengan

$$w_{hi}(x) = \frac{\frac{1}{h}K\left(\frac{x - x_i}{h}\right)}{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} = \frac{K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

4. Pemilihan *Bandwidth* Optimal

Bandwidth (*h*) adalah parameter pemulus (*smoothing*) yang berfungsi untuk mengontrol kemulusan dari kurva yang diestimasi. *Bandwidth* yang terlalu kecil akan menghasilkan kurva yang *under-smoothing* yaitu sangat kasar dan sangat fluktuatif, dan sebaliknya *bandwidth* yang terlalu lebar akan menghasilkan kurva yang *over-smoothing*yaitu sangat mulus, tetapi tidak sesuai dengan pola data (Hardle, 1994). Oleh karena itu perlu dipilih *bandwidth* yang optimal. Salah satu metode untuk mendapatkan *h* optimal adalah dengan menggunakan kriteria *Generalized Cross Validation* (GCV), yang didefinisikan sebagai berikut:

$$GCV = \frac{MSE}{\left(\frac{1}{2}tr(I = H(h))\right)^2}$$

dengan $MSE = \frac{1}{n}\sum_{i=1}^{n}(y - m_h(x_i))^2$. Kebaikan suatu estimator dapat dilihat dari tingkat kesalahannya. Semakin kecil tingkat kesalahannya semakin baik estimasinya. Kriteria untuk menentukan estimator terbaik dalam model regresi nonprametrik, antara lain :

1. *Mean Square Error* (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{y}_i)^2$$

2. *Root Mean Square Error* (RMSE)

$$RMSE = \sqrt{MSE}$$

3. *Mean Absolute Deviation* (MAD)

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

METODE

1. Sumber Data

Data yang kami gunakan dalam penelitian ini adalah IPM tahun 2020 dari 34 provinsi yang merupakan data sekunder yang diperoleh dari BPS sebagai unit observasi.

2. Varibel Penelitian

Variabel yang digunakan dalam penelitian ini terdiri dari variabel respon (Y) dan Variabel prediktor (X).

Y : Indeks pembangunan manusia tahun 2020,

X : Rata-rata lama sekolah (RLS).

3. Analisis Data

Tahapan analisis yang digunakan dalam penelitian ini adalah :

a. Membuat analisis statistika deskriptif.
b. Menguji asumsi klasik.
c. Melihat pola sebaran data yang akan digunakan dalam penelitian dengan cara membuat scatterplot pada setiap variabel.
d. Menentukan fungsi kernel yang digunakan, yaitu Kernel Gaussian

$$K(u) = \frac{1}{\sqrt{2\pi}}\exp\left(\frac{1}{2}(-u^2)\right), -\infty < x < \infty.$$

e. Melakukan pemilihan *bandwidth* (h) yang optimal dan memodelkan nilai *bandwidth* (h) optimal dengan metode *Generalized Cross Validation* (GCV) yang minimun

$$GCV(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{m}_h(x_i)}{n^{-1}tr[\boldsymbol{I} - \boldsymbol{A}(\mathrm{h})]}\right)^2$$

Nilai *bandwidth* optimal dapat dicari dengan menggunakan rumus berikut:

$$h_{opt} = 1.06an^{-\frac{1}{5}}$$

dengan

$$a = \min\left\{s, \frac{R}{1{,}34}\right\}$$
$$n = banyak\ data$$
$$R = jangkauan\ kuartil$$
$$s = standar\ deviasi$$

f. Mengestimasi dengan estimator *Nadaraya-Watson*

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)}$$

g. Mengestimasi model nonparametrik

$$y_i = f(x_i) + \varepsilon_i, i = 1,2, \dots, n.$$

Keterangan

y_i : variabel prediktor ke – i
$f(x_i)$: fungsi nonparametrik ke – i yang tidak diketahui
ε_i : error ke – i yang diasumsikan menyebar N~(0;σ^2)

h. Mengambil kesimpulan yang di dapatkan dari hasil penilitan

ANALISIS DAN PEMBAHASAN

1. Statistika Deskriptif Data

Tabel 1. Hasil Deskriptif Data

Variable	Mean	StDev	Varian	Min	Max	Median
IPM	71,08	3,90	15,23	60,44	80,77	71,43
RLS	8,65	0,93	0,86	6,69	11,13	8,71

Berdasarkan hasil deskriptif diatas terlihat bahwa Indeks Pembangunan Manusia dari 34 provinsi di Indonesia terendah sebesar 60,44 dan tertinggi sebesar 80,77. Rata-rata lama sekolah (RLS) dari 34 provinsi di Indonesia terendah 6.69 dan tertinggi sebesar 11,13.

2. Uji Asumsi Klasik

a. Kenormalan

Uji kenormalan menggunakan Uji Anderson Darling, dengan hasil sebagai berikut:

Tabel 2. Hasil Uji Anderson Darling

Asumsi	p-value
Kenormalan	0,608

Berdasarkan hasil uji kenormalan dengan Uji Anderson-Darling terlihat bahwa data berdistribusi normal ,ini ditunjukkan oleh nilai *p-value* > (0,05).

Tabel 3. Hasil Uji Durbin Watson

Asumsi	p-value
Autokorelasi	0,764

Berdasarkan hasil uji autokorelasi dengan Uji Durbin Watson terlihat bahwa data tidak terjadi autokorelasi. Hal ini ditunjukkan oleh nilai *p-value* > (0,05).

b. Heteroskedasitas

Pengujian Heteroskedasitas dalam penelitian ini menggunakan Uji Glejser dengan hasil sebagai berikut :

Tabel 4. Hasil Uji Glejser

Asumsi	p-value
Heteroskedasitas	0,824

Berdasarkan hasil Uji Glejser terlihat bahwa data tidak terjadi Heteroskedasitas. Hal ini dibuktikan dengan *p-value* > (0,05).

c. Multikolinearitas

Pengujian Multikolinearitas dalam penelitian ini menggunakan Uji V*arince Inflation Factor* (VIF), dengan hasil sebagai berikut :

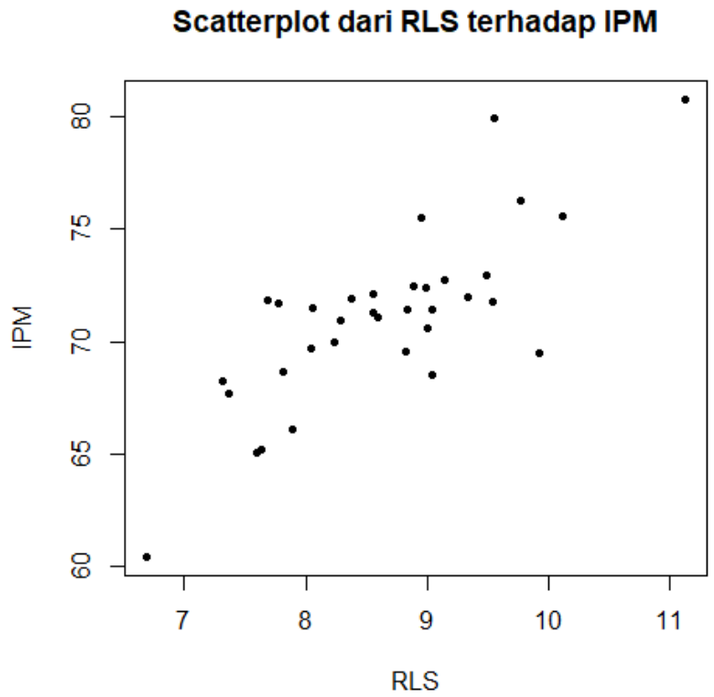
Tabel 5. Hasil Uji VIF

Asumsi	p-value
Multikolinearitas	1,000

Tabel 5 menunjukkan nilai VIF setiap variabel predictor dalam penelitian ini. Terlihat bahwa setiap variabel memiliki nilai VIF yang lebih kecil dari 10, maka dapat disimpulkan bahwa tidak terjadi multikolinearitas dalam data penelitian ini.

3. Regresi Non Parametrik Kernel

a. Identifikasi Pola Sebaran Data



Gambar 1. Scatterplot dari variabel rata-rata lama sekolah terhadap indek pembangunan manusia.

Berdasarkan gambar 1 menunjukkan data variabel prediktor yaitu rata-rata lama sekolah terhadap variabel respon yaitu indeks pembangunan manusia menyebar secara acak atau tidak membentuk suatu garis lurus (linear) maupun suatu lengkungan (non linear). Sehingga dapat disimpulkan bahwa data variabel ini merupakan komponen nonparametrik.

b. Estimasi Model Regresi Nonparametrik

Dengan nilai U adalah :

$$u = \frac{x - x_i}{h} = \frac{X_i - X_j}{h} \; dengan \; i, j = 1,2,3, \dots ,34$$

Diperoleh fungsi kernel yang sesuai untuk penelitian ini adalah Fungsi Kernel Gaussian sebagai berikut:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2}\left(\frac{X_j - X_i}{h}\right)^2\right)$$

Dengan menggunakan variabel nonparametrik yaitu rata-rata lama sekolah, estimasi diperoleh dengan mensubstitusikan fungsi kernel ke estimator *Nadaraya-Watson*, sebagai berikut

$$\widehat{m}(x) = \frac{\sum_{i=1}^{34}\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X_j - X_i}{h}\right)^2\right) Y_i\right)}{\sum_{i=1}^{34} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{X_j - X_i}{h}\right)^2\right)} + \varepsilon_i$$

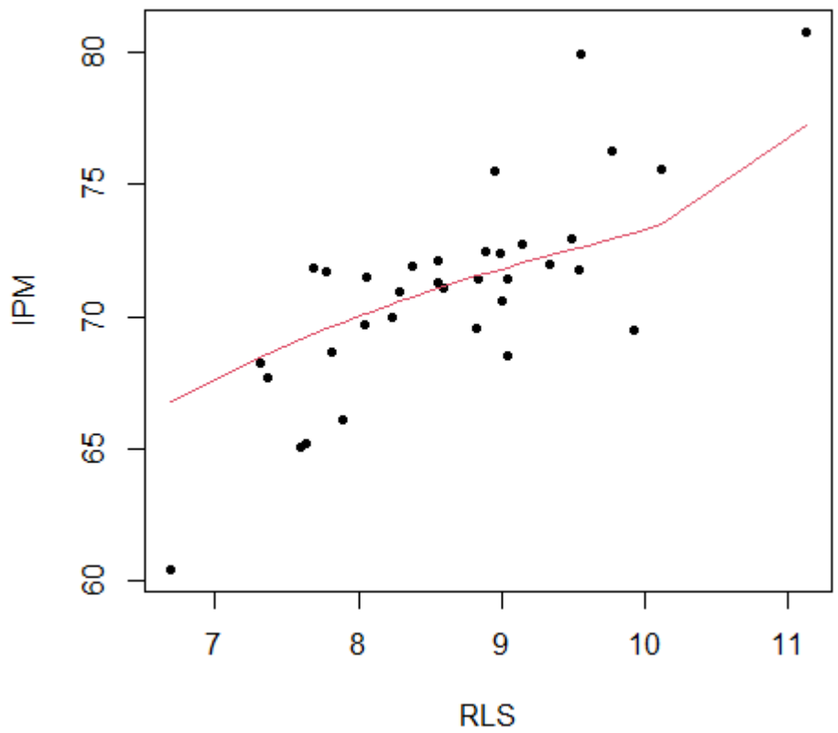
Sehingga didapatkan model berikut

$$\hat{y} = \widehat{m}(x) + \varepsilon_i$$

$$\hat{y} = \frac{\sum_{i=1}^{34}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_j - X_i}{h}\right)^2\right)Y_i\right)}{\sum_{i=1}^{34}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_j - X_i}{h}\right)^2\right)} + \varepsilon_i$$

4. Pemilihan *Bandwidth* Optimal

Metode GCV merupakan metode yang digunakan dalam menentkan pemilihan *bandwidth* optimal, Nilai awal dari penentuan *bandwidth* optimal sebagai acuan dalam estimasi model regresi nonparametrik. Diperoleh nilai GCV minimal sebesar 6.850771, dengan nilai *bandwidth* (h) sebesar 0.6936735. Grafik hasil pemulusan sebagai berikut



Gambar 2. Grafik pemulusan dengan h = 0.6936735.

Berikut persamaan model indeks pembangunan manusia dengan pendekatan nonparametrik kernel *Nadaraya-Watson* dengan metode pemilihan *bandwidth* optimal GCV sebagai berikut:

$$\hat{y} = \frac{\sum_{i=1}^{34}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_1 - X_{1i}}{0.6936735}\right)^2\right)Y_i\right)}{\sum_{i=1}^{34}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_1 - X_{1i}}{0.6936735}\right)^2\right)} + \varepsilon_i$$

KESIMPULAN

Berdasarkan hasil penelitian, maka dapat disimpulkan bahwa estimasi model regresi nonparametrik dengan menggunakan estimator *Nadaraya-Watson* Dengan nilai *bandwidth* optimal sebesar 0.6936735 pada kasus indeks pembangunan manusia adalah sebagai berikut:

$$\hat{y} = \frac{\sum_{i=1}^{34}\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_1 - X_{1i}}{0.6936735}\right)^2\right)Y_i\right)}{\sum_{i=1}^{34}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{X_1 - X_{1i}}{0.6936735}\right)^2\right)} + \varepsilon_i$$

Dalam penelitian ini, permasalahan yang dikaji masih sangat terbatas dikarenakan diperlukan penilitan lebih lanjut pada estimasi model indeks pembangunan manusia berdasarkan rata-rata lama sekolah dapat digunakan faktor lain yang mempengaruhi indeks pembangunan manusia, misalnya harapan lama sekolah, angka harapan hidup, pengeluaran perkapita, dan lainnya.

DAFTAR PUSTAKA

Alfiani, Mifta Luthfin, *et al.* 2014. *Model Regresi Nonparametrik Berdasarkan Estimator Polinomial Lokal Kernel pada Kasus Pertumbuhan Balita.* Jurnal Statistika. Vol. 2 No.1. Universitas Muhammadiyah Semarang.

Aydin, Dursun. 2007. *A Comparison of the Nonparametric Regres- sion Models using Smoothing Spline and Kernel Regression*. World Academy of Science, Engineering and Technology, 36, 253-257, Turkey. http://www.waset.org/journals/waset/v36/v36-46.pdf. Diakses tanggal 9 Februari 2010.

Eubank, R. 1998. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker.New York.

Hardle, W. 1994. *Applied Nonparametric Regression*. Cambridge University Press.New York.

Lestari, B., & Budiantara, I. N. (2010). *"Spline Estimator of Triple Response Nonparametric Regression Model"*. Jurnal Ilmu Dasar, Vol. 11, hal. 17-22.

Maharani, Agni Horti, *el al*. 2012. *Pemodelan Berat Badan Balita dengan Menggunakan Regresi Kernel*. Jurnal Matematika. Vol. 4 No. 3. Universitas Andalas.

Razak, Resti Anita, *et al*. 2019, *Penerapan Cross Validation (CV) dalam Pemilihan Bandwidth Optimal pada Pemodelan Regresi Nonparametrik Kernel (Studi Kasus: Gizi Buruk pada Balita di Indonesia),* Prosiding, Universitas Muhammadiyah Semarang.

Wand M.P. and M.C.Jones. 1995. *Kernel Smoothing*. Chapman and Hall. New York.