

BAB I

PENDAHULUAN

1.1 Latar Belakang

Analisis regresi adalah salah satu teknik analisis statistika yang digunakan untuk mengetahui bentuk hubungan antara dua variabel atau lebih. Variabel-variabel dalam regresi sendiri terbagi atas dua jenis, yaitu variabel respon atau variabel dependen dan variabel prediktor atau variabel independen. Dalam penyelesaian suatu analisis regresi, dapat dilakukan dengan dua model pendekatan, yaitu model pendekatan regresi parametrik dan model pendekatan regresi nonparametrik (Putri *et al.*, 2022). Model pendekatan parametrik digunakan apabila terdapat informasi terkait pola datanya, baik itu berpola garis lurus, kuadratik, atau lainnya. Selain itu, model parametrik sangat terikat dengan asumsi-asumsi tertentu yang harus terpenuhi yaitu *error* harus bersifat independen dan berdistribusi normal identik (Suparti *et al.*, 2018).

Sebaliknya, model pendekatan nonparametrik digunakan ketika bentuk pola data tidak diketahui dan jenis pendekatan ini tidak terikat dengan asumsi-asumsi tertentu yang harus terpenuhi, sehingga memberikan fleksibilitas yang lebih besar (Abdy, 2019). Dalam beberapa kasus, terdapat data observasi yang akan dikaji tidak selalu membentuk suatu pola tertentu seperti garis lurus, sehingga pendekatan regresi parametrik kurang tepat digunakan dalam memodelkan data tersebut (Suparti *et al.*, 2018).

Sebagai alternatif, dilakukan pendekatan dengan model nonparametrik. Walaupun pendekatan nonparametrik digunakan untuk mengatasi pemodelan data yang tidak memenuhi suatu pola atau asumsi tertentu, akan tetapi pendekatan ini

dapat digunakan dalam memodelkan data dalam bentuk pola acak ataupun data yang tidak memiliki suatu pola tertentu (Suparti *et al.*, 2018). Estimasi fungsi dalam model nonparametrik dilakukan berdasarkan data observasi dengan memanfaatkan teknik pemulusan atau *smoothing*. Terdapat beberapa teknik *smoothing* dalam model nonparametrik yaitu kernel, spline, polinomial lokal, deret fourier, dan wavelet (Eubank, 1999).

Regresi kernel adalah teknik statistika yang digunakan untuk menemukan hubungan antara variabel prediktor dan variabel respon dengan menggunakan bobot fungsi densitas kernel. Regresi kernel memiliki kemampuan yang baik dalam memodelkan data yang memiliki bentuk pola secara acak, dan dapat mencapai tingkat konvergen yang relatif cepat (Hardle, 1994). Regresi kernel sendiri terbagi dua berdasarkan jenis variabelnya, yaitu regresi kernel univariabel dan multivariabel. Regresi kernel univariabel meliputi satu variabel respon dan satu variabel prediktor sedangkan regresi kernel multivariabel meliputi satu variabel respon dan memiliki variabel prediktor lebih dari satu.

Pemodelan regresi kernel menggunakan bobot fungsi densitas kernel untuk mendapatkan nilai estimasi dari fungsi regresi kernel. Menurut Halim dan Bisono (2006), estimator fungsi regresi kernel terbagi menjadi tiga jenis, yaitu estimator *Nadaraya-Watson*, estimator *Priestley-Chao*, dan estimator *Gasser-Muller*. Estimator fungsi regresi yang umum digunakan pada regresi kernel adalah estimator *Nadaraya-Watson* (Lamusu *et al.*, 2020). Selain itu, estimator *Nadaraya-Watson* memiliki sifat yang konsisten secara asimtotik. Artinya, ketika data bertambah, hasil estimasi yang dihasilkan juga akan semakin mendekati nilai observasi. Estimator ini juga fleksibel karena tidak memerlukan parameter yang

kompleks, sehingga sesuai digunakan dalam kasus yang melibatkan lebih dari satu variabel prediktor atau multivariabel (Tosatto *et al.*, 2020).

Selain itu, untuk mendapatkan estimasi yang akurat, penting untuk memahami peran dari estimator densitas kernel dalam menentukan fungsi densitas suatu data tanpa mengasumsikan bentuk distribusi tertentu (Hardle, 1994). Estimator densitas kernel pada fungsi regresi bergantung pada dua parameter yaitu fungsi kernel dan *bandwith*. Terdapat beberapa jenis fungsi kernel, antara lain Gaussian, *Triangle*, Epanechnikov, *Biweight*, dan *Uniform* (Ogden, 1997). Fungsi kernel *triangle* merupakan fungsi bobot yang memiliki keunggulan dalam perhitungan yang lebih mudah dan cepat (Sukarsa & Srinadi, 2012). Namun, di antara kedua faktor tersebut, pemilihan *bandwith* memiliki pengaruh yang signifikan dibandingkan dengan fungsi kernel. Penentuan *bandwith* atau parameter pemulus yang tepat menjadi bagian yang penting untuk memastikan tidak terjadinya *overfitting* ataupun *underfitting* pada model (Hardle, 1994).

Penentuan *bandwith* yang optimal berguna untuk mendapatkan model yang terbaik. Metode yang umum digunakan untuk menentukan *bandwith* optimal adalah metode optimasi *Mean Square Error* (MSE), *Cross-Validation* (CV) dan *Generalized Cross-Validation* (GCV). Metode ini bertujuan untuk meminimalkan kesalahan estimasi dengan memilih nilai *bandwith* yang menghasilkan nilai prediksi terbaik berdasarkan kriteria yang ditetapkan oleh masing-masing optimasi (Hardle, 1994). Optimisasi dengan metode MSE sering kali menyebabkan semakin kecil nilai *bandwith* yang digunakan, semakin kecil pula nilai MSE yang dihasilkan sehingga dapat menghasilkan model yang *undersmooth*. Oleh karena itu, metode

CV dan GCV dipilih karena lebih efektif dalam menghindari model yang *undersmooth* dan memberikan estimasi yang lebih akurat (Suparti *et al.*, 2018).

Berbagai penelitian telah dilakukan terkait dengan penggunaan regresi nonparametrik kernel. Sadek dan Mohammed (2024) melakukan evaluasi terhadap dua jenis pemodelan regresi yaitu regresi parametrik dan regresi nonparametrik terhadap data yang memiliki pola yang acak. Hasil penelitian menunjukkan bahwa dengan menggunakan regresi nonparametrik kernel didapatkan hasil koefisien determinasi yang tinggi yaitu sebesar 95% dibandingkan dengan regresi parametrik yang memiliki nilai koefisien determinasi sebesar 23%. Puspitasari *et al.* (2012) melakukan analisis terkait data saham dengan menggunakan regresi parametrik dan regresi nonparametrik, dimana model terbaiknya dihasilkan oleh regresi nonparametrik kernel. Dalam analisis tersebut, terdapat perbandingan berbagai fungsi kernel yang diujikan. Model terbaik yang diperoleh dengan regresi nonparametrik kernel menggunakan fungsi kernel *triangle* yang menghasilkan nilai MSE yang paling kecil yaitu sebesar 6987,787.

Selain itu, penelitian lain yang dilakukan oleh Astuti *et al.* (2018) menemukan bahwa dengan menggunakan berbagai jenis fungsi kernel, kurva regresi yang dihasilkan tidak jauh berbeda satu sama lain, menunjukkan konsistensi dalam hasil estimasi. Penelitian lain dilakukan oleh Razak *et al.* (2019) menggunakan regresi nonparametrik kernel multivariabel untuk menentukan model terbaik terkait data gizi buruk pada balita di Indonesia. Dengan optimasi CV, penelitian tersebut menghasilkan koefisien determinasi yang kuat yaitu sebesar 84,73%, mengindikasikan model yang baik. Lamusu *et al.* (2021) juga mengaplikasikan regresi nonparametrik kernel untuk mengestimasi produksi jagung, dan hasilnya

menunjukkan bahwa model terbaik diperoleh dengan optimasi GCV, yang memberikan estimasi yang lebih tinggi dibandingkan dengan CV yang digunakan pada data tersebut dengan ukuran signifikansi *R-squared* yang tinggi.

Berdasarkan uraian yang telah dipaparkan, penelitian ini melakukan pemodelan regresi nonparametrik kernel multivariabel dengan menggunakan fungsi kernel *triangle* dan membandingkan metode optimasi CV dan GCV. Selain itu, peneliti terdahulu dalam pengolahan data masih menggunakan interaksi utama pada program R yaitu *Command Line Interface* (CLI). Saat ini dalam R terdapat pengembangan dalam interaksinya, salah satunya adalah *Graphical User Interface* (GUI) yang berguna untuk mempermudah melakukan pengolahan data yang memiliki iterasi yang banyak. Dalam penelitian ini, dilakukan pengembangan program GUI untuk memudahkan perhitungan dan meningkatkan efisiensi dalam mencari model terbaik. Dengan bantuan GUI, perhitungan berbagai kombinasi *bandwith* dapat dilakukan dengan lebih cepat dan optimal sehingga menghasilkan model yang terbaik. Hal ini berbeda dengan pendekatan manual yang diterapkan dalam melakukan pengujian berbagai *bandwith* pada penelitian sebelumnya.

Sebagai implementasinya, dilakukan analisis data terkait kualitas pendidikan terhadap Indeks Pembangunan Manusia. Menurut *United Nations Development Program* (UNDP), IPM adalah ukuran perbandingan yang mencakup dimensi harapan hidup, pendidikan dan juga standar hidup untuk semua negara di seluruh dunia (Raghuvanshi & Verma, 2024). IPM atau biasa disebut dengan *Human Development Index* (HDI) digunakan sebagai pengklasifikasian apakah suatu negara dapat dikatakan sebagai negara maju, berkembang atau keterbelakang (Rahayu & Akbarita, 2022).

Berdasarkan *United Nations Development Programme*, indikator yang mengukur kualitas pendidikan dalam IPM adalah rata-rata lama sekolah dan harapan lama sekolah. Dimensi pendidikan ini mencerminkan tingkat melek huruf dan tingkat pendidikan dalam suatu populasi (Rahayu & Akbarita, 2022). Penelitian ini berfokus pada dimensi pendidikan karena pendidikan mempunyai peranan penting dalam meningkatkan kualitas hidup dan berkontribusi langsung pada perkembangan ekonomi dan sosial masyarakat. Data pada dimensi pendidikan juga sering kali bersifat fluktuatif karena berbagai faktor sosial dan ekonomi (Sabrina *et al.*, 2022). Menurut Badan Pusat Statistik, nilai IPM di Indonesia pada tahun 2020 hingga 2023 menunjukkan peningkatan sebesar 0,72 persen per tahun, dari 72,81 pada 2020 menjadi 74,39 pada tahun 2023. Kedua indikator pembentuk dimensi pendidikan juga menunjukkan peningkatan dari periode 2020 hingga 2023. Rata-rata indikator harapan lama sekolah meningkat 0,43 persen per tahun, dari 12,98 pada 2020 menjadi 13,15 pada tahun 2023. Rata-rata lama sekolah meningkat 1,13 persen per tahun, dari 8,48 pada 2020 menjadi 8,77 pada tahun 2023.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, maka ditentukan beberapa rumusan masalah, yaitu:

1. Bagaimana pemodelan pada data kualitas pendidikan terhadap IPM di Indonesia pada Tahun 2023 dengan menggunakan optimasi CV dan GCV?
2. Bagaimana perbandingan kebaikan model dan evaluasi kinerja model antara optimasi CV dan GCV dalam pemilihan *bandwith* optimal pada regresi nonparametrik kernel multivariabel pada data kualitas pendidikan terhadap IPM di Indonesia Tahun 2023?

3. Bagaimana pengembangan R GUI yang dirancang untuk melakukan pemodelan regresi nonparametrik kernel multivariabel dengan optimasi CV dan GCV?

1.3 Batasan Masalah

Penentuan batasan masalah diperlukan dalam penelitian ini guna agar penelitian lebih fokus dan tidak meluas dari pembahasan yang dimaksud. Dalam penelitian ini, penulis membatasinya pada ruang lingkup penelitian sebagai berikut:

1. Analisis yang digunakan adalah analisis regresi nonparametrik kernel dengan menggunakan estimator kernel *Nadaraya-Watson* dengan memanfaatkan fungsi kernel *triangle* sebagai fungsi pembobotnya dengan rentang *bandwith* yang diujikan adalah 1 sampai 2 dengan kenaikan *bandwith* 0,1.
2. Data yang digunakan dalam penelitian ini yaitu data Indeks Pembangunan Manusia (IPM) yang memengaruhi data rata-rata lama sekolah dan harapan lama sekolah yang berada dalam cakupan dimensi pendidikan berdasarkan provinsi di Indonesia pada Tahun 2023. Data tersebut akan dibagi menjadi data *training* dan data *testing* guna untuk menentukan model terbaiknya. Data *training* dan data *testing* memiliki perbandingan sebesar 80:20 yang dibagi secara acak.
3. Pada program R GUI, variabel prediktor terbatas hingga dua variabel prediktor dan satu variabel respon.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah diuraikan, tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Memperoleh model terbaik pada data kualitas pendidikan terhadap IPM di Indonesia pada Tahun 2023 dengan menggunakan optimasi CV dan GCV.

2. Membandingkan kebaikan model dan evaluasi kinerja model yang dihasilkan antar kedua optimasi yaitu CV dan GCV pada data kualitas pendidikan terhadap IPM di Indonesia pada Tahun 2023.
3. Menghasilkan GUI dalam program R untuk memudahkan penentuan model terbaik dengan menggunakan regresi nonparametrik kernel multivariabel dengan optimasi CV dan GCV.