BAB 3
METODOLOGI PENELITIAN

Bab ini membahas mengenai metode penelitian dan metode *Boosted Regression Tree* (BRT) yang digunakan untuk mengetahui besar pengaruh pada masing-masing peubah penjelas terhadap peubah responnya.

3.1. Prosedur Penelitian

Metode yang digunakan dan dibahas pada penelitian ini adalah metode BRT. Penelitian ini dilakukan dengan melakukan studi literatur yakni dengan mencari referensi teori yang berkaitan dengan metode BRT yang tersedia di berbagai sumber contohnya jurnal, buku, dan internet. Selanjutnya megkonstruksi algoritma BRT dengan pengaplikasiannya pada suatu data lalu menyajikannya dalam bahasa pemrograman R.

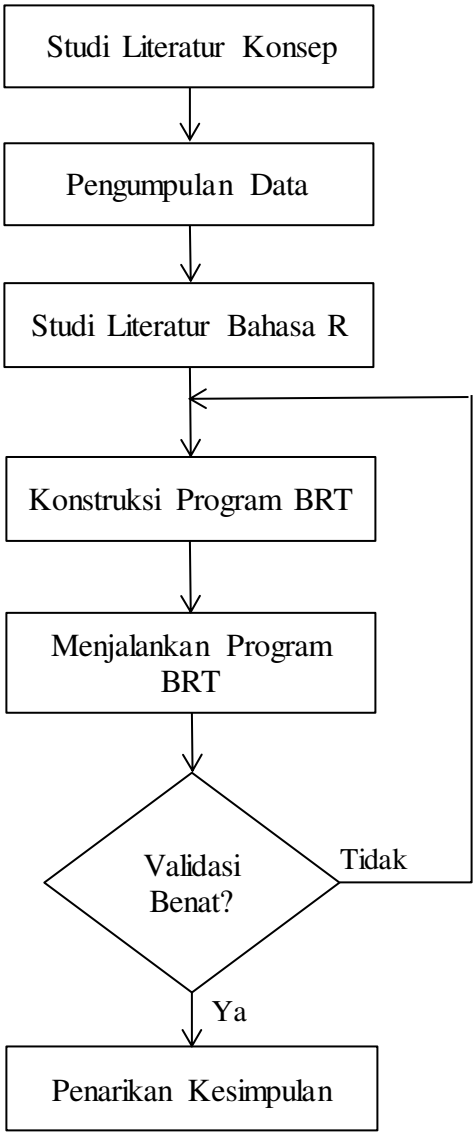
Data yang digunakan pada penelitian ini adalah data sekunder yang diperoleh dari beberapa publikasi Badan Pusat Statistik (BPS) di Provinsi Jawa Timur tahun 2019 tersedia di *jatim.bps.go.id*. Data yang digunakan pada penelitian ini adalah data Jumlah Tindak Pidana yang Dilaporkan Menurut Kabupaten/Kota sebagai peubah responnya dan sebelas peubah penjelasnya dengan banyaknya objek pada data sebanyak 38 objek yakni Kabupaten/Kota yang berada di Provinsi Jawa Timur.

Langkah-langkah yang dilakukan pada penelitian ini adalah sebagai berikut:

1. Melakukan studi literatur mengenai konsep dasar mengenai pohon regresi, *gradient boosting*, yang selanjutnya dikonstruksi menjadi BRT.
2. Mengumpulkan data dari berbagai publikasi BPS yang nantinya akan digunakan untuk pengaplikasian metode BRT.
3. Melakukan studi literatur mengenai bahasa pemrograman R.
4. Mengkonstruksi program aplikasi untuk mengetahui pengaruh peubah penjelas menggunakan metode BRT dengan bahasa pemrograman R.

5. Melakukan proses penggunaan BRT.
6. Melakukan proses validasi yaitu dengan membandingkan hasil yang diperoleh dari yang sudah dibentuk dengan *packages* yang sudah tersedia dalam aplikasi R.
7. Menarik kesimpulan yang berkaitan dengan tujuan penelitian.

Langkah-langkah pada metodologi penelitian yang telah dijelaskan di atas dapat disajikan dalam bentuk diagram alur seperti tersaji pada gambar 3.1.



Gambar 3.1 Diagram Alur Metodologi Penelitian

3.2. Boosted Regression Trees (BRT)

Boosted Regression Trees (BRT) merupakan salah satu metode yang menggabungkan dua algoritma yaitu algoritma pohon regresi dan *boosting* (Hastie, Leathwick, & Elith, 2008). Metode *boosting* yang digunakan pada penelitian ini adalah metode *gradient boosting*. Metode *gradient boosting* memperbaiki pohon dengan mengacu pada fungsi kerugiannya.

Pada bab sebelumnya telah dibahas mengenai materi *gradient boosting*. Pada persamaan (2.9) merupakan hasil pendekatan menggunakan "*greedy stagewise*". Untuk m=1,2,…,M pada asumsi parameter $\{\beta_m, \mathbf{a_m}\}_1^M$ sebagai titik data terdekat dengan data uji berhingga $\{y_i, \mathbf{x_i}\}_1^N$. Namun, masalah optimasi pada persamaan (2.9) berpontensi sulit untuk diselesaikan satu per satu dengan metode kuadrat terkecil pada persamaan (2.11).

Optimasi parameter tunggal atau ρ_m pada persamaan (2.12) berdasarkan kriteria kerugian (ψ). Pada kasus BRT $h(x;\ \mathbf{a}_m)$ atau *base learner* berupa pohon regresi.

Pseudoresponse adalah penurunan *gradient* yang curam guna membantu optimasi parameter. *Pseudoresponse* yang digunakan pada metode BRT sama seperti pada *gradient boosting* yaitu:

$$\tilde{y}_i = -\mathrm{g}_m(\mathbf{x}_i) = -\left[\frac{\partial \psi\big(y_i, F(\mathbf{x}_i)\big)}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}. \qquad (3.1)$$

dimana $\psi\big(y_i, F(\mathbf{x}_i)\big)$ adalah fungsi kerugian dan $F_{m-1}(\mathbf{x})$ adalah nilai prediksi pada pohon $m-1$

Gradient boosting untuk pohon regresi mengkhususkan pendekatan untuk kasus dimana *base learner* $h(\mathbf{x};\mathbf{a})$ adalah terminal simpul L pada pohon regresi. Pada setiap m iterasi, pohon regresi mempartisi ruang **x** menjadi daerah L yang saing lepas (*disjoint*) $\{R_{lm}\}_{l=1}^L$ dan memprediksi nilai konstanta terpisah pada masing masing daerahnya.

$$h(\mathbf{x}; \{R_{lm}\}_{l=1}^L) = \sum_{i-1}^{N} \bar{y}_{lm} 1(\mathbf{x} \in R_{lm}) \qquad (3.2)$$

dimana $\bar{y}_{lm}= mean_{\mathbf{x}_i \in R_{lm}}(\tilde{y}_{im})$ adalah rata-rata dari persamaan (3.1) di setiap wilayah R_{lm}. Parameter dari *base learner* ini adalah variabel pemisah dan titik pemisah yang sesuai untuk mendefinisikan pohon, kemudian menentukan daerah $\{R_{lm}\}_{l=1}^{L}$ yang berkoresponden dengan partisi pada iterasi ke-m. Dengan pohon regresi, persamaan (2.12) dapat diselesaikan secara terpisah dalam setiap region R_{lm} yang ditentukan oleh simpul terminal l yang sesuai dengan pohon ke-m. Karena pohon (3.2) memprediksikan nilai konstan $\bar{y}_{lm}$dalam setiap region R_{lm}, sehingga solusi untuk persamaan (2.12) direduksi menjadi estimasi lokal sederhana berdasarkan kriteria ψ yang bersesuaian dengan simpul terminal l dan pohon ke-m.

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{x_i \in R_{lm}} \psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma) \tag{3.3}$$

dengan

ψ = fungsi kerugian

R_{lm} = region dengan simpul terminal l dan pohon ke-m

$F_{m-1}(\mathbf{x}_i)$ = nilai prediksi pada pohon $m-1$

y_i = nilai observasi ke-I

γ_{lm} = *line search* pada simpul terminal l dan pohon ke-m.

Perkiraan untuk $F_{m-1}(\mathbf{x}_i)$ kemudian secara terpisah diperbarui pada masing-masing region yang berkresponden

$$F_m = F_{m-1}(\mathbf{x}_i) + \nu.\gamma_m 1(\mathbf{x} \in R_{lm}) \tag{3.4}$$

dengan ν adalah parameter "*shrinkage*" memiliki nilai $0 < \nu \leq 1$ untuk mengontrol *learning rate* dari prosedur, secara empiris (Friedman, 1999) menemukan bahwa nilai kecil untuk nilai ν $(\nu \leq 1)$ menyebabkan kesalahan generalisasi menjadi lebih baik.

Sehingga dari uraian diatas mengarah kepada algoritma untuk menggeneralisir boosting untuk pohon regresi adalah sebagai berikut:

$F_0(\mathbf{x}) = \arg\min_{\rho} \sum_{i=1}^{N} L(y_i, \psi).$

$For\ m = 1\ to\ M\ do$

$$\tilde{y}_i = -\left[\frac{\partial \psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N$$

$$\{R_{lm}\}_{l=1}^{L} = L - terminal\ node\ tree(\{\tilde{y}_{im}, x_i\}_1^N)$$

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{x_i \in R_{lm}} \psi(y_i, F_{m-1}(x_i) + \gamma)$$

$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu.\gamma_m 1(x \in R_{lm})$$

EndFor

3.3. Perancangan Program Aplikasi

Subbab ini membahas mengenai perancangan program aplikasi yaitu rancangan data masukan, data keluaran, algoritma dari program aplikasi BRT dengan bantuan bahasa pemrograman RStudio.

3.3.1.Data Masukan

Data masukan yang dimuat dalam program BRT disajikan dalam Tabel 3.1

Tabel 3.1 Daftar Data Masukan

Data	Nama Variabel	Tipe Data
Jumlah Tindak Pidana	Y	Numerik
Pembagian Wilayah Administratif	X_1	String
Kepadatan Penduduk	X_2	Numerik
Jarak ke Ibukota Surabaya	X_3	String
PDRB per Kapita atas Dasar Harga	X_4	Numerik
Tingkat Pengangguran Terbuka	X_5	Numerik
Persentase Penduduk Miskin	X_6	Numerik
Jumlah Pemuda	X_7	Numerik
Angka Partisipasi Kasar SD	X_8	Numerik
Angka Partisipasi Kasar SMP	X_9	Numerik
Angka Partisipasi Kasar SMA	X_{10}	Numerik
Kemantapan Jalan	X_{11}	Numerik

3.3.2.Data Keluaran

Data keluaran yang akan diperoleh dari hasil pemrograman BRT adalah sebagai berikut seperti tersaji pada tabel 3.2

Tabel 3.2 Daftar Data Keluaran

Data	Nama Variabel	Tipe Data
RMSE	rmse_fit4	Numerik
Pohon Terbaik	perf_gbm	String
Relative Influence	rel_inf	Numerik

3.3.3.Algoritma Pemrograman

Perancangan program aplikasi untuk mengetahui berapa besar berapa pengaruh masing-masing peubah penjelas terhadap peubah responnya menggunakan metode BRT (*Boosted Regression Tree*) dilakukan dengan bahasa pemrograman R. Dengan program aplikasi R akan dibuat fungsi yang dibutuhkan untuk mengolah data agar diperoleh hasil yang diinginkan, dalam penelitian ini, hasil yang diinginkan adalah hasil dari penggunaan metode BRT.

Tipe data peubah penjelasnya dapat berupa ordinal dan nominal. Khusus pada penelitian ini terdapat 11 peubah penjelas yang dianggap sebagai faktor yang memPengaruhi peubah responnya yaitu tingkat kriminalitas. Selanjutnya, data dari 11 peubah penjelasnya diproses oleh fungsi yang sudah terbentuk, ~~maka~~ akan diperoleh nilai-nilai pada peubah penjelasnya. Nilai yang dihasilkan merupakan besar pengaruh terhadap peubah responnya.

Program aplikasi BRT dengan bahasa pemrograman RStudio adalah sebagai berikut:

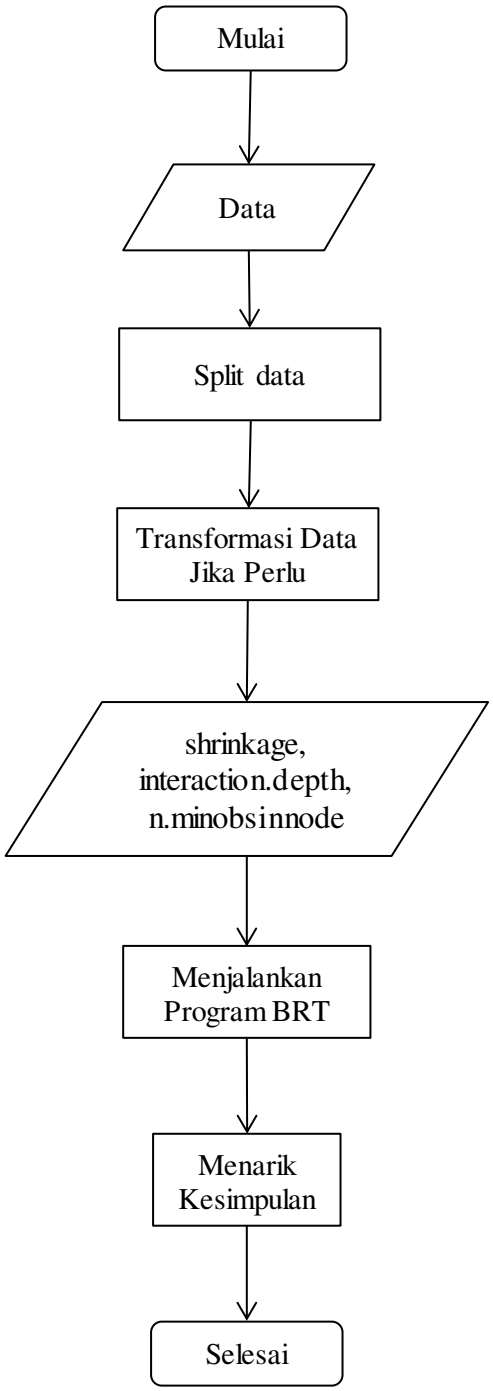
1. Menginput data yang akan digunakan.

   Pada penelitian ini data yang digunakan adalah jumlah tindak pidana sebagai peubah respon, sisanya seperti pembangian wilayah administratif, kepadatan penduduk, jarak wilayah yang digunakan ke ibukota Surabaya, PDRB per kapita atas dasar harga, tingkat pengangguran terbuka, persentase penduduk miskin, jumlah pemuda, kemantapan jalan, angka partisipasi kasar SD, angka partisipasi kasar SMP, dan angka partisipasi SMA.
2. Melakukan perubahan tipe data, maksudnya yaitu apabila terdapat data bertipe string, maka data tersebut diubah data tersebut diubah tipe datanya menjadi tipe nominal agar memudahkan proses.
3. Pemisahan data menjadi dua bagian secara acak, yaitu data uji sebanyak 30% dan data latih awal 70%. Pembentukan pohon regresi dilakukan pada

data latih atau training data, kemudian data uji atau test data digunakan untuk validasi pohon regresi yang terbentuk.

4. Melakukan proses penggunaan metode BRT, beberapa hal yang harus diperhatikan sebelum menggunakan metode BRT adalah:
   a. Nilai shrinkage atau learning rate untuk penyusutan tiap pohon yang terbentuk setelahnya. Semakin kecil learning rate maka semakin teliti dan banyak pohon yang terbentuk untuk mencapai optimal sehingga mampu mereduksi overfitting namun, proses boosting akan berlangsung lama. Sebaliknya, Semakin besar learning rate semakin cepat proses boosting terhenti tetapi hasil yang diperoleh kemungkinan masih memungkinkan terjadinya overfitting. Pada penelitian kali ini akan diambil nilai nilai shrinkagenya adalah 0.01.
   b. Nilai interaction.depth yakni banyaknya split atau pemisahan maksimal yang dilakukan pada setiap pohon yang terbentuk. Pemberian nilai pada penelitian ini adalah dengan melihat pohon terbaik dengan menggunakan pohon regersi. Setelah terbentuknya pohon regresi menghasilkan nilai interaction.depth sebesar 3.
   c. Nilai n.minobsinnode yakni nilai yang menyatakan jumlah minimum pengamatan pada setiap simpul terminal. Menurut Breiman et.al (1993) banyaknya amatan pada simpul akhir ≤ 5 maka proses penyekatan rekursif berakhir. Maka pada penelitian kali ini diambil nilai n.minobsinnodenya adalah 5.
5. Melakukan analisis BRT dengan langsung dengan memanggil *function* yang sudah dibuat. Di dalamnya terdiri dari pengaruh relative (*relative influence*), plot pengaruh antar dua peubah penjelas terhadap peubah responnya, dan pohon optimal yang terbentuk.
6. Menarik kesimpulan dari analisis yang telah dilakukan, seper.

Langkah-langkah program aplikasi BRT yang telah dijelaskan di atas dapat disajikan dalam bentuk diagram alur seperti tersaji pada gambar 3.2.



Gambar 3.2 Diagram Alur Program BRT

Kemudian setelah dikonstruksi program BRT maka untuk melihat bahwa pendekatan BRT adalah memperbaiki metode pohon regresi adalah dengan membandingkan nilai RMSE kedua hasil antara pohon regresi dan BRT.