BAB I
PENDAHULUAN

1.1 Latar Belakang

Data merupakan gabungan informasi yang didapat dari suatu observasi, baik berupa angka, simbol, kalimat, dan lainnya. Perkembangan data saat ini sudah meningkat pesat dalam kehidupan manusia. Berbagai bidang ilmu pengetahuan sering menggunakan analisis data untuk memperoleh informasi yang ada di dalam sekumpulan data dan memanfaatkannya dalam memecahkan suatu masalah. Dari sekian banyaknya, analisis data yang umum sering dipakai, yaitu analisis regresi (Bazdaric dkk., 2021). Analisis regresi kini berkembang dan banyak digunakan dalam berbagai bidang penelitian, seperti pendidikan, kesehatan, bisnis, dan bidang studi lainnya (Iqbal, 2020).

Analisis regresi merupakan teknik yang dipakai dalam memprediksi pola hubungan variabel bebas dengan variabel terikat, sehingga diperoleh suatu bentuk persamaan yang menyatakan hubungan fungsional antarvariabel secara matematis (Sudjana, 2005). Secara umum, variabel terikat pada analisis regresi bersifat kontinu, namun terkadang terdapat variabel terikat yang bersifat kategorik. Salah satu jenis regresi yang dapat digunakan dalam menganalisis variabel terikatnya bersifat kategorik yaitu regresi logistik.

Regresi logistik digunakan untuk melihat keterkaitan antara variabel terikat kategorikal dengan satu atau lebih variabel bebas kategorikal atau kontinu (Hosmer dkk., 2013). Berdasarkan jenis skala data variabel terikatnya regresi logistik dapat dikelompokkan menjadi biner, multinomial, dan ordinal. Hubungan antara variabel bebas kategorikal atau kontinu dengan variabel terikatnya memiliki dua kategori saja dapat dianalisis dengan menggunakan regresi logistik biner. Adapun hubungan antarvariabel ini dapat dilihat berdasarkan pemodelan yang terbentuk dari penaksir koefisien regresi.

Metode yang biasa digunakan untuk menentukan nilai penaksir koefisien regresi, yaitu *Ordinary Least Square* (OLS). OLS adalah metode yang dipakai untuk memperkirakan nilai koefisien regresi linear dengan meminimalkan jumlah

kuadrat sisaan. OLS dapat digunakan ketika banyaknya amatan melebihi banyaknya variabel bebas dan matriks bersifat nonsingular (Setiawan & Sutikno, 2010). Sebaliknya, OLS tidak dapat diimplementasikan pada data yang bersifat singular seperti data berdimensi tinggi.

Data berdimensi tinggi merupakan data dengan karakteristik jumlah variabel bebas lebih banyak dibandingkan amatan (Narisetty, 2020). Sering kali terdapat beberapa permasalahan yang timbul dalam analisis data berdimensi tinggi. Dalam menduga koefisien regresi, metode OLS tidak dapat digunakan dalam data berdimensi tinggi. Selain itu, masalah yang sering muncul pada data berdimensi tinggi, yaitu adanya permasalahan multikolinearitas (Zhao dkk., 2020). Multikolinearitas menyebabkan varians parameter yang ditaksir terlalu besar dan mengurangi akurasi penaksiran (Vatcheva dkk., 2016). Padahal salah satu asumsi regresi adalah antarvariabel bebas yang dimasukkan dalam model tidak ada multikolinearitas, sehingga data seperti ini membutuhkan pendekatan yang berbeda dalam analisis data. Terdapat berbagai metode untuk menangani analisis regresi pada data berdimensi tinggi. Salah satunya dengan cara menyeleksi variabel bebas melalui metode LASSO.

LASSO digunakan untuk menyeleksi variabel dan mengecilkan beberapa koefisien menuju nol (Hastie dkk., 2009). Karena data berdimensi tinggi memiliki variabel bebas yang banyak, sangat memungkinkan variabel satu dengan yang lainnya memiliki kemiripan dan membentuk suatu kelompok variabel. LASSO memiliki keterbatasan dalam mengatasi kelompok variabel, sehingga dikembangkan metode yang dapat penyeleksian terhadap kelompok variabel bebas, yaitu *Group* LASSO.

Group LASSO sering kali digunakan dalam pemilihan variabel pada data variabel bebas yang membentuk suatu kelompok (Chen & Xiang, 2017). *Group* LASSO dapat digunakan dalam pemilihan variabel dan mengatasi multikolinearitas serta dapat digunakan dalam data kategorik (El Sheikh dkk., 2021). Pada penelitian sebelumnya, Yunus dkk. (2017) mengkaji *Group* LASSO dalam data berkorelasi tinggi, hasilnya menunjukkan bahwa *Group* LASSO lebih baik dari LASSO dan *Least Square*. Chen dkk. (2020) juga menggunakan *Group*

LASSO dan LASSO dalam menganalisis variabel yang berkaitan dengan kasus TBC di Jawa Barat. Hasilnya menunjukkan bahwa *Group* LASSO lebih baik daripada LASSO dalam pemilihan variabel.

Indeks Pembangunan Manusia (IPM) merupakan indikator penting untuk mengukur capaian pembangunan kualitas hidup manusia. IPM juga dimanfaatkan dalam melihat tingkatan pembangunan suatu wilayah atau negara (Sapaat dkk., 2020). IPM diukur dengan tiga indikator, yaitu dimensi pendidikan, dimensi kesehatan, dan dimensi ekonomi. Ketiga indikator tersebut tidak berdiri sendiri melainkan terdapat banyak faktor pendukung yang menunjang indikator tersebut.

Menurut Badan Pusat Statistik (BPS), capaian IPM Provinsi Jawa Barat tahun 2020 mencapai pada angka 72,09 dengan rincian dari 27 kota/kabupaten sebanyak 13 kota/kabupaten IPM berkategori sedang, 11 kota/kabupaten berkategori tinggi, dan 3 kota/kabupaten berkategori sangat tinggi (Badan Pusat Statistik, 2021). Dari capaian IPM Provinsi Jawa Barat tersebut, sebanyak 10 kota/kabupaten berada di atas capaian IPM Jawa Barat dan sisanya berada di bawah capaian IPM Jawa Barat. Adanya ketimpangan capaian IPM antarkota/kabupaten inilah yang menjadi permasalahan, sehingga perlu diidentifikasi lebih lanjut agar permasalahan tersebut bisa teratasi.

Pada umumnya, dalam mengidentifikasi faktor-faktor yang mempengaruhi terhadap capaian IPM dianalisis menggunakan analisis regresi. Apabila bentuk datanya gabungan dari *cross section* dan *time series*, maka regresi data panel dapat digunakan. Penelitian dengan metode tersebut pernah dilakukan oleh Digdowiseiso (2021) dan Suryani (2021). Sedangkan, apabila bentuk datanya hanya berupa data *cross section* dan variabel terikatnya berbentuk kategori, umumnya digunakan regresi logistik. Selain digunakan untuk memprediksi pola hubungan variabel bebas terhadap variabel terikat, regresi logistik dapat digunakan dalam klasifikasi (Edgar & Manz, 2017). Oleh karena itu, regresi logistik dapat digunakan juga dalam memetakan capaian IPM antarkota/kabupaten.

Dalam beberapa tahun terakhir, terdapat penelitian yang mengkaji terkait faktor-faktor yang berpengaruh terhadap capaian IPM dengan menggunakan

model regresi logistik. Putra dan Ratnasari (2015) menganalisis faktor-faktor yang mempengaruhi IPM Provinsi Jawa Timur menggunakan 13 variabel bebas yang dibagi menjadi 3 kelompok dengan metode regresi logistik *ridge*. Hasil penelitian tersebut menghasilkan 5 variabel bebas yang berpengaruh dengan ketepatan klasifikasi model sebesar 97,36%. Sari dan Purhadi (2021) meneliti faktor IPM tiga provinsi di Pulau Jawa tahun 2019 menggunakan regresi logistik dengan 5 variabel bebas. Hasil penelitian tersebut menunjukkan faktor yang signifikan terhadap IPM Provinsi Jawa Barat, yaitu tingkat kemiskinan dan sumber air bersih dengan ketepatan klasifikasi model 77,78%.

Khairunnisa dkk. (2022) juga menganalisis faktor-faktor yang berpengaruh terhadap IPM Provinsi Jawa Barat pada tahun 2020 menggunakan 3 variabel bebas dengan metode regresi logistik biner. Hasilnya menunjukkan Persentase Penduduk Miskin berpengaruh terhadap IPM Provinsi Jawa Barat pada tahun 2020. Akan tetapi, variabel yang digunakan dan diduga berpengaruh dalam penelitian-penelitian sebelumnya masih sangat sedikit, khususnya penelitian di daerah Jawa Barat. Oleh karena itu, perlu adanya penelitian baru dengan menambah variabel yang diperkirakan berpengaruh terhadap IPM Provinsi Jawa Barat sehingga hasil yang diperoleh bisa lebih komprehensif.

Penambahan banyak variabel bebas yang besar tentunya akan merubah karakteristik data tersebut menjadi data berdimensi tinggi. Penelitian dengan karakteristik data tersebut pernah dikaji oleh Sunandi (2021) pada kasus analisis faktor yang berpengaruh terhadap IPM Provinsi Bengkulu tahun 2019. Pada penelitiannya, Sunandi membandingkan antara metode LASSO dan *Group* LASSO. Penelitian tersebut menggunakan simulasi 36 variabel bebas yang dibagi menjadi 12 kelompok dengan amatan sebanyak 30 buah. Hasilnya menunjukkan *Group* LASSO lebih baik daripada LASSO dalam penanganan data berdimensi tinggi pada kasus tersebut.

Penelitian ini menggunakan data capaian IPM Kota/Kabupaten di Jawa Barat tahun 2020 dengan 27 amatan dan 40 variabel bebas yang terbagi menjadi 6 kelompok. Dalam penelitian ini, variabel terikat dikategorikan menjadi dua kategori, sehingga untuk mengetahui faktor-faktor yang mempengaruhi capaian

IPM dikembangkan dengan regresi logistik biner. Data tersebut tergolong kedalam data berdimensi tinggi, karena banyak variabel bebas lebih besar daripada amatannya. Apabila dianalisis dengan menggunakan regresi logistik biner biasa akan menimbulkan permasalahan seperti adanya multikolinearitas. Oleh karena itu, penelitian ini menggunakan regresi logistik biner dengan metode *Group* LASSO untuk mengidentifikasi faktor-faktor yang mempengaruhi capaian IPM Kota/Kabupaten di Jawa Barat tahun 2020 dalam mengatasi permasalahan data berdimensi tinggi.

1.2 Rumusan Masalah

Berdasarkan penjelasan latar belakang tersebut, maka didapat rumusan masalah seperti berikut.

1. Bagaimana model regresi logistik biner dengan *Group* LASSO untuk mengetahui faktor-faktor yang mempengaruhi capaian IPM Kota dan Kabupaten di Jawa Barat tahun 2020?
2. Faktor-faktor manakah yang paling dominan mempengaruhi capaian IPM Kota dan Kabupaten di Jawa Barat tahun 2020?

1.3 Tujuan Penelitian

Sejalan dengan rumusan masalah tersebut, tujuan penelitian ini antara lain seperti berikut.

1. Memperoleh model regresi logistik biner menggunakan metode *Group* LASSO pada data capaian IPM Kota dan Kabupaten di Jawa Barat tahun 2020.
2. Memperoleh faktor-faktor yang paling dominan mempengaruhi IPM Kota dan Kabupaten di Jawa Barat tahun 2020.

1.4 Manfaat Penelitian

Adapun manfaat yang diperoleh dari penulisan penelitian ini antara lain seperti berikut.

1. Secara teoritis, hasil dari penelitian ini dapat memberi wawasan baru kepada para pembaca berkaitan dengan analisis regresi logistik biner menggunakan metode *Group* LASSO.
2. Secara praktis, hasil penelitian ini dapat menghasilkan faktor-faktor yang paling dominan mempengaruhi capaian IPM Kota dan Kabupaten di Jawa Barat tahun 2020, sehingga bisa dipelajari untuk meningkatkan pembangunan dalam periode ke depan.