Jurnal Ilmiah Sosio Agribis (JISA) ISSN: 1412-1816 (p), ISSN: 2614-4549 (e)
Volume 22, Nomor 2 : 123-132

PENERAPAN METODE REGRESI *LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR (LASSO)* DAN REGRESI LINIER UNTUK MEMPREDIKSI TINGKAT KEMISKINAN DI INDONESIA

Application of Least Absolute Shrinkage and Selection Operator (LASSO) Methods and Linear Regression to Predict Poverty Levels in Indonesia

Eka Rossalina Fitria[1*], Fatchur Rozci[2]
[1*,2,3] Department Agribusiness, Faculty of Agriculture,
Universitas Pembangunan Nasional "Veteran" Jawa Timur
East Java, Indonesia
*Correspondence author: Eka Rossalina Fitria
Email: ekarossalina30@gmail.com

ABSTRACT

Poverty is defined as a situation where a person or family is unable to meet the basic needs for survival, such as clothing, food, shelter, and education. This study aims to compare the accuracy of the Least Absolute Shrinkage and Selection Operator (Lasso) regression method and linear regression in predicting poverty levels in Indonesia and to choose the best model from the two methods used. The results of this study indicate that of the two algorithms used, namely the linear regression algorithm and Lasso regression, the algorithm that has a higher level of accuracy for predicting poverty rates in each province in Indonesia is the linear regression algorithm because it has a lower MSE value and has the value of R^2 is closer to 1 than the Lasso regression algorithm. In addition, the results of the analysis show that the variables that have the highest influence on poverty rates in provinces in Indonesia are education, as well as the Human Development Index (IPM).
Keywords: *Poverty Rate Prediction, Lasso Regression, Linear Regression*

ABSTRAK

Kemiskinan diartikan sebagai suatu keadaan dimana seseorang atau keluarga tidak mampu untuk memenuhi kebutuhan dasar untuk bertahan hidup, seperti sandang, pangan, papan, dan pendidikan. Studi ini bertujuan untuk membandingkan tingkat akurasi metode regresi *Least Absolute Shrinkage and Selection Operator (LASSO)* dan regresi linier dalam memprediksi tingkat kemiskinan di Indonesia serta memilih model terbaik dari kedua metode yang digunakan tersebut. Hasil dari studi ini menunjukkan bahwa dari kedua algoritma yang digunakan yaitu algoritma regresi linier dan regresi Lasso, algoritma yang memiliki tingkat akurasi lebih tinggi untuk memprediksi tingkat kemiskinan pada masing-masing provinsi di Indonesia adalah algoritma regresi linier karena memiliki nilai MSE yang lebih rendah serta memiliki nilai R^2 mendekati 1 dibandingkan algoritma regresi Lasso. Selain itu, dari hasil analisis menunjukkan bahwa variabel yang memiliki pengaruh tertinggi terhadap tingkat kemiskinan pada provinsi di Indonesia yaitu pendidikan, serta Indeks Pembangunan Manusia (IPM).
Kata Kunci : Prediksi Tingkat Kemiskinan, Regresi Lasso, Regresi Linier

PENDAHULUAN

Kemiskinan merupakan salah satu masalah yang masih dihadapi oleh semua negara berkembang sehingga perlu segera diselesaikan. Kemiskinan bukanlah satu entitas tunggal yang muncul dengan sendirinya, namun didasarkan pada berbagai faktor yang saling berinteraksi dan menimbulkan konflik sosial-ekonomi yang berkelanjutan. Seseorang atau sekelompok orang dapat dikatakan miskin apabila dirinya tidak mampu mencapai tingkat kemakmuran ekonomi yang dianggap sebagai persyaratan minimum untuk standar hidup tertentu. Badan Pusat Statistik (BPS) menyatakan bahwa kemiskinan adalah ketidakmampuan seseorang untuk memenuhi

kebutuhan hidup minimum, baik makanan maupun non makanan. Seseorang dianggap miskin jika dirinya hanya mampu memenuhi kebutuhan pangannya kurang dari 2.100 kalori per orang per hari (dari 52 jenis komoditi yang dianggap mewakili pola konsumsi penduduk kelas menengah ke bawah) dan konsumsi atas non makanan (dari 45 jenis komoditi menurut kesepakatan di tingkat nasional dan tidak dibedakan antara wilayah pedesaan dan perkotaan). Tolok ukur kecukupan 2.100 kalori berlaku untuk semua kelompok umur, jenis kelamin, dan perkiraan aktivitas fisik, berat badan, dan perkiraan status fisiologis penduduk. Ukuran ini sering disebut sebagai garis kemiskinan. Penduduk dengan pendapatan di bawah garis kemiskinan dianggap miskin (Statistik, 2011).

Pandangan ekonomi baru menganggap tujuan utama pembangunan ekonomi bukan hanya pertumbuhan PDB semata, tapi juga pengentasan kemiskinan, penanggulangan ketimpangan pendapatan dan penyediaan lapangan kerja dalam konteks perekonomian yang terus berkembang (Todaro & Smith, 2004). Hal tersebut dapat dimaknai bahwa kemiskinan menjadi salah satu masalah yang harus diatasi. Usaha pemerintah pusat maupun daerah dalam pengentasan masalah kemiskinan sangatlah serius, bahkan merupakan salah satu program prioritas. Berbagai program bantuan telah diusahakan pemerintah untuk disalurkan kepada para penduduk miskin, seperti Bantuan Langsung Tunai (BLT), Program Keluarga Harapan (PKH), dan lain-lain, namun tidak semua kebijakan dan program yang dilaksanakan menunjukkan hasil yang optimal. Masih terjadi kesenjangan antara rencana dengan pencapaian tujuan karena kebijakan dan program pengentasan kemiskinan lebih berorientasi pada program sektoral. Selain itu, salah satu kesulitan yang seringkali dihadapi oleh pemerintah adalah proses pembagian bantuan sosial yang tidak merata dan tepat sasaran. Oleh karena itu, diperlukan suatu strategi pengentasan kemiskinan yang terpadu, terintegrasi, dan sinergis sehingga dapat menyelesaikan masalah secara tuntas (Permana & Arianti, 2012).

Tabel 1. Jumlah dan Persentase Penduduk Miskin di Indonesia (Mar 2017-Sep 2021)

No	Periode	Jumlah jiwa	Persentase (%)
1	Q1 2017	27.771.220	10,64
2	Q3 2017	26.582.990	10,12
3	Q1 2018	25.949.800	9,82
4	Q3 2018	25.674.580	9,66
5	Q1 2019	25.144.720	9,41
6	Q3 2019	24.785.870	9,22
7	Q1 2020	26.424.020	9,78
8	Q3 2020	27.549.690	10,19
9	Q1 2021	27.542.770	10,14
10	Q3 2021	26.503.650	9,71

Sumber: (Statistik, 2022)

Berdasarkan data Badan Pusat Statistik (BPS) pada tahun 2022, tercatat bahwa jumlah penduduk miskin di Indonesia semakin membaik tiap tahunnya, yang ditunjukkan dengan adanya penurunan yang cukup progresif antara kedua periode, yaitu periode bulan Maret dan bulan September di setiap tahun. Jumlah penduduk miskin di Indonesia berkurang sebanyak 1,04 jiwa menjadi 26,5 juta pada September 2021 dibandingkan Maret 2021. Sedangkan jika dibandingkan dengan bulan September 2020, jumlah penduduk miskin juga berkurang 1,05 juta jiwa. Demikian pula dengan persentase penduduk miskin juga turun 0,43 persen poin menjadi 9,71% pada September 2021 dibanding Maret 2021. Jika dibanding September 2020, angka kemiskinan juga turun 0,48 persen poin. Meskipun menurun, tetapi persentase penduduk miskin tersebut masih lebih tinggi dibanding posisi sebelum terjadi pandemi Covid-19 (Statistik, 2022).

Sumber: (Statistik, 2022)

Gambar 1.
Angka Melek Huruf Penduduk Pedesaan di Indonesia Tahun 2017-2021

Membaca dan menulis merupakan kemampuan keaksaraan dasar yang harus dikuasai oleh setiap individu. Pengertian angka melek huruf (AMH) menurut BPS adalah proporsi penduduk usia 15 tahun ke atas yang mempunyai kemampuan membaca dan menulis huruf latin dan huruf lainnya, tanpa harus mengerti apa yang dibaca atau ditulisnya terhadap penduduk usia 15 tahun ke atas. Menurut data Badan Pusat Statistik (BPS), angka melek huruf (AMH) penduduk pedesaan usia 15 tahun keatas Indonesia terus mengalami peningkatan sejak satu dekade terakhir. AMH penduduk pedesaan usia 15 tahun ke atas nasional tercatat sebesar 93,65% pada 2021 atau naik tipis sebesar 0,01 poin dari tahun sebelumnya sebesar 93,64%. Hal ini menunjukkan bahwa sebagian besar penduduk berusia 15 tahun ke atas di Indonesia telah melek huruf atau telah memiliki kemampuan membaca dan menulis huruf latin dan huruf lainnya. Namun terlepas dari persentase angka melek huruf yang cukup besar tersebut, penduduk Indonesia belum bisa dikatakan melek huruf secara keseluruhan, dikarenakan masih terdapat sekitar 7% penduduk Indonesia pada usia 15 tahun ke atas di daerah pedesaan yang belum memiliki kemampuan membaca dan menulis huruf latin. Hal tersebut dikarenakan masih kurang meratanya fasilitas pendidikan yang dapat dengan mudah dijangkau oleh masyarakat pedesaan dengan kemampuan ekonomi menengah ke bawah, atau kurang sadarnya masyarakat tersebut terhadap pentingnya memiliki kemampuan membaca dan menulis.

Faktor-faktor yang telah dijabarkan di atas adalah faktor-faktor yang diduga mempengaruhi tingkat kemiskinan di Indonesia. Untuk itu, perlu dilakukan pengujian terhadap faktor-faktor yang mempengaruhi tingkat kemiskinan lalu diklasifikasikan. Salah satu pemodelan yang dapat digunakan untuk klasifikasi tingkat kemiskinan tersebut adalah dengan menggunakan data mining. *Data mining* merupakan proses untuk menemukan pengetahuan (*knowledge discovery*) yang diambil dari sekumpulan data yang volumenya besar. *Data mining* menggunakan data yang sudah ada dan relevan dari tahun-tahun sebelumnya, lalu membuat beberapa model untuk mengidentifikasi pola-pola diantara atribut-atribut yang ada di dalam dataset. Model adalah penyajian matematis (persamaan linear sederhana dan/atau persamaan kompleks yang sangat tidak linear) yang mengidentifikasi pola-pola diantara berbagai atribut objek yang ada di dalam dataset. Dalam *data mining* terdapat beberapa fungsi yaitu *data mining* untuk aturan asosiasi (*Association Rules*), *data mining* untuk klasifikasi (*Classification*), *data mining* untuk pengelompokan (*Clustering*), *data mining* untuk prediksi (*prediction/forecasting*). Penelitian ini bertujuan untuk memanfaatkan data mining untuk memprediksi/meramalkan bagaimana tingkat kemiskinan di Indonesia di masa mendatang. Berdasarkan klasifikasi model yang dihasilkan nantinya, kita dapat memilih model terbaik untuk memprediksi bagaimana tingkat kemiskinan

Indonesia di masa depan. Dalam penelitian ini, penulis ingin mencoba untuk membandingkan model algoritma regresi Lasso dan regresi linier.

Dalam penyusunan Penerapan Metode Regresi *Least Absolute Shrinkage and Selection Operator* (LASSO) dan Regresi Linier untuk Memprediksi Tingkat Kemiskinan di Indonesia sebagai gambaran bagi pemerintah di setiap provinsi di Indonesia bahwa masih terdapat banyak faktor yang mempengaruhi tingkat kemiskinan, sehingga pemerintah dapat lebih memperhatikan daerah-daerah pedesaan yang masih belum maju baik secara ekonomi maupun infrastruktur, agar lebih diperhatikan lagi sehingga tingkat kemiskinan di Indonesia dapat membaik di setiap tahunnya. Mengetahui metode regresi manakah yang dapat memprediksi tingkat kemiskinan di Indonesia dilihat dari nilai akurasinya. Mengetahui faktor-faktor mana sajakah yang mempengaruhi tingkat kemiskinan di Indonesia.

Tinjauan pustaka yang relevan ialah kemiskinan diartikan ketidakmampuan seseorang dilihat dari segi ekonomi untuk memenuhi kebutuhan dasar makanan dan bukan makanan yang diukur dari pengeluaran. Penduduk miskin adalah penduduk yang rata-rata pengeluaran perkapita perbulan dibawah garis kemiskinan, baik itu pengeluaran untuk makanan atau non makanan. Seseorang dikatakan miskin apabila hidupnya serba kekurangan, sehingga tidak mampu memenuhi kebutuhannya. Kemiskinan didefinisikan sebagai kondisi dimana seseorang mengalami kekurangan uang dan barang untuk menjamin keberlangsungan hidup. Garis kemiskinan absolut sangat penting untuk menilai efek dari kebijakan anti kemiskinan antar waktu atau memperkirakan dampak dari suatu proyek terhadap kemiskinan (Statistik, 2018). BAPPEDA (2011) kemiskinan dapat disebabkan oleh kelangkaan alat pemenuh kebutuhan dasar, ataupun sulitnya akses terhadap pendidikan dan pekerjaan. Oleh karena itu, tingkat kemiskinan dapat mempengaruhi nilai IPM.

Kemiskinan lazim digambarkan sebagai gejala kekurangan pendapatan untuk memenuhi kebutuhan pokok untuk bertahan hidup. Sekelompok masyarakat dikatakan dibawah garis kemiskinan jika pendapatan kelompok masyarakat tidak cukup memenuhi kebutuhan hidup yang paling pokok seperti pangan, pakaian, dan tempat tinggal (Setiadi & Kolip, 2011). Dalam arti luas, kemiskinan merupakan suatu fenomena multiface atau multidimensional (Suryawati, 2005). Hidup dalam kemiskinan bukan hanya hidup dalam kekurangan uang dan tingkat pendapatan rendah, tetapi juga banyak hal lain, seperti: tingkat kesehatan, pendidikan rendah, perlakuan tidak adil dalam hukum, kerentanan terhadap ancaman tindak kriminal, ketidakberdayaan menghadapi kekuasaan, dan ketidakberdayaan dalam menentukan jalan hidupnya sendiri. Kemiskinan dapat dibagi dalam empat bentuk, yaitu:

1. Kemiskinan absolut: apabila penghasilan di bawah garis kemiskinan atau tidak cukup untuk memenuhi kebutuhan pangan, pakaian, Kesehatan, tempat tinggal, dan pendidikan yang diperlukan untuk bisa hidup dan bekerja.
2. Kemiskinan relatif: kondisi kemiskinan yang dikarenakan adanya pengaruh kebijakan pembangunan yang belum menjangkau seluruh lapisan masyarakat, sehingga terjadi ketimpangan pada penghasilan.
3. Kemiskinan kultural: kondisi kemiskinan yang mengacu pada persoalan sikap individu atau masyarakat yang disebabkan oleh faktor budaya, seperti tidak mau berusaha meningkatkan taraf kehidupan, malas, pemboros, tidak kreatif, bahkan ketika bantuan dari luar sudah tersedia.
4. Kemiskinan struktural: kondisi kemiskinan yang disebabkan karena terbatasnya akses terhadap sumber daya yang terjadi dalam suatu sistem sosial-budaya dan sosial-politik yang tidak mendukung pengentasan kemiskinan, tetapi seringkali membiarkan kemiskinan semakin subur (Suryawati, 2005).

Terdapat sembilan kriteria yang menunjukkan bahwa seseorang dapat dikatakan miskin: Ketidakmampuan dalam memenuhi kebutuhan dasar seperti sandang, pangan, dan papan. Cacat fisik atau mental yang menyebabkan seseorang terganggu atau bahkan tidak dapat bekerja. Lingkungan yang buruk, seperti kurang beruntung sejak lahir (anak terlantar, korban wanita KDRT, janda miskin, dan kelompok terpencil). Kualitas SDM yang rendah, seperti buta huruf,

pendidikan yang rendah, kondisi yang tidak sehat, serta keterbatasan SDA (tanah yang kurang subur, lokasi tidak memadai, listrik, air, dan infrastruktur terbatas). Berpenghasilan rendah, serta terbatasnya fasilitas umum yang dapat menunjang perekonomian. Kondisi lingkungan yang buruk menyebabkan kurangnya lapangan pekerjaan yang kurang luas. Keterbatasan akses terhadap kebutuhan dasar seperti air bersih, pendidikan, sanitasi, dan transportasi. Tidak ada jaminan masa depan (karena tidak adanya investasi untuk pendidikan keluarga dan perlindungan sosial dari negara dan masyarakat). Tidak terlibat dalam kegiatan sosial masyarakat (Suharto, 2009).

Pendidikan merupakan salah satu unsur ilmu pengetahuan, sikap dan keterampilan yang berperilaku biasanya dapat berada di lingkungan sekolah atau pendidikan formal. Namun tidak hanya pendidikan formal, melalui pendidikan individu memiliki kemampuan untuk mengembangkan diri guna mencapai penghidupan yang lebih baik, baik menurut jenjang pendidikan formal maupun informal, yang tercermin dari angka melek huruf. Angka melek huruf juga dapat menjadi indikator perkembangan pendidikan penduduk. Semakin tinggi literasi atau angka melek huruf, maka semakin tinggi kualitas dan mutu sumber daya manusia. Diasumsikan bahwa orang yang dapat membaca dan menulis memiliki keterampilan dan kemampuan karena dapat menyerap informasi baik secara lisan maupun tulisan (Statistik, 2011).

Tingkat pendidikan diukur dengan dua indikator, yaitu angka melek huruf dan rata-rata lama sekolah. Angka melek huruf adalah persentase penduduk berusia 15 tahun ke atas yang mampu membaca dan menulis huruf latin atau huruf lainnya. Bobot indikator ini adalah dua pertiga. Sepertiga sisa pembobotan diberikan kepada rata-rata tahun sekolah (MYS/*Mean Year of Schooling*), yaitu rata-rata jumlah tahun yang pernah dijalani pada semua jenjang pendidikan formal bagi penduduk berusia 15 tahun ke atas. Indikator ini dihitung dari variabel pendidikan tertinggi yang ditamatkan dan tingkat pendidikan yang sedang dijalani.

IPM pertama kali diperkenalkan oleh *United Nations Development Programme* (UNDP) pada tahun 1990. Badan Pusat Statistik (BPS) mengganti beberapa indikator yang digunakan dalam perhitungan IPM, diantaranya Angka Melek Huruf yang diubah menjadi Angka Harapan Lama Sekolah, PDRB per kapita yang diganti dengan Produk Nasional Bruto (PNB) per kapita. IPM berperan sebagai indikator sekaligus alat ukur pencapaian kualitas hidup suatu masyarakat. Sebagai alat ukur, IPM ditinjau melalui pendekatan tiga dimensi dasar, yaitu umur panjang dan hidup sehat, pengetahuan dan standar hidup layak. Setiap dimensi tersebut diwakili oleh beberapa indikator, diantaranya indikator kesehatan, pendidikan dan pengeluaran per kapita (Statistik, 2018).

Indeks Pembangunan Manusia (IPM) atau *Human Development Index* (HDI) mencakup tiga bidang yaitu usia hidup (*longevity*), pengetahuan (*knowledge*), dan standar hidup layak (*decent living*). Indeks Pembangunan Manusia (IPM) merupakan salah satu pendekatan untuk mengukur tingkat keberhasilan pembangunan manusia. Meskipun tidak dapat mengukur semua dimensi dari pembangunan, namun mampu IPM mampu mengukur dimensi pokok pembangunan manusia yang dinilai, dapat mencerminkan kemampuan dasar Individu. Adapun manfaat dari IPM adalah sebagai berikut:

a. Mengukur keberhasilan dalam upaya membangun kualitas hidup manusia (masyarakat).
b. Menentukan peringkat atau level pembangunan suatu wilayah/negara.
c. Bagi Indonesia, IPM merupakan data strategis karena selain sebagai ukuran kinerja pemerintah, IPM juga sebagai salah satu alokator penentuan Dana Alokasi Umum (DAU) (BPS, 2013).

Data mining merupakan rangkaian proses yang mengkaji nilai tambah berupa informasi yang sebelumnya tidak diketahui dari suatu dataset (Pramudiono, 2007). *Data mining* sering juga disebut dengan *knowledge discovery in database* (KDD). KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola, atau hubungan dalam dataset berukuran besar (Santosa, 2007). Sejalan dengan pendapat dari yang mengatakan bahwa *data mining* diartikan sebagai proses menemukan pola dalam data. Proses ini otomatis atau seringkali semi otomatis. Pola yang ditemukan harus signifikan dan pola tersebut menawarkan

keuntungan, biasanya keuntungan finansial. Data yang dibutuhkan dalam jumlah besar (Witten & Frank, 2002). *Data mining* menjadi dua kategori utama (Han & Kamber, 2006), yaitu:

a. Prediktif: Tujuan dari tugas prediktif adalah untuk memprediksi nilai dari atribut tertentu berdasarkan pada nilai atribut lainnya. Atribut yang diprediksi sering disebut sebagai target atau variabel dependen, sedangkan atribut yang digunakan untuk membuat prediksi dikenal sebagai variabel penjelas atau independen.
b. Deskriptif: Tujuan dari tugas deskriptif adalah untuk memperoleh pola (korelasi, tren, cluster, wilayah, dan anomali) yang meringkas hubungan paling penting dalam data. Tugas penambangan data deskriptif seringkali bersifat eksplorasi dan seringkali membutuhkan teknik post-processing untuk memvalidasi dan menjelaskan hasilnya.

METODE PENELITIAN

Penelitian ini merupakan jenis penelitian deskriptif kuantitatif dengan memaparkan fenomena yang ada di tengah masyarakat serta menggunakan angka-angka untuk menjabarkan karakteristik permasalahan dan hasil penelitian yang akan dipaparkan. Data yang digunakan dalam penelitian ini adalah data sekunder dengan metode pengumpulan data menggunakan *Library Research* (penelitian kepustakaan), dimana peneliti melakukan studi kepustakaan dari berbagai literatur memperoleh informasi atau peralatan dasar yang berkaitan dengan penelitian, seperti jurnal-jurnal, buletin-buletin, laporan, serta bahan bacaan lainnya yang berkaitan dengan masalah yang diteliti. Dataset diperoleh dari Badan Pusat Statistik (BPS).

Metode pengumpulan data adalah suatu cara yang digunakan untuk memperoleh data suatu objek yang kemudian digunakan untuk menyusun hasil penelitian. Jenis data yang digunakan dalam penelitian ini adalah data sekunder. Data sekunder merupakan metode pengumpulan data yang tidak langsung memberikan data pada pengumpul data, misalnya lewat orang lain atau lewat dokumen.frekuensi data yang digunakan dalam penelitian ini adalah data runtut waktu (*time series*). Jadi, metode pengumpulan data yang digunakan meliputi metode dokumentasi yang merupakan catatan peristiwa yang sudah berlalu, berbentuk tulisan, gambar, atau karya – karya monumental dari seseorang dan metode studi kepustakaan yakni cara pengumpulan data yang mempelajari buku–buku, jurnal, tesis, skripsi, dan literatur lainnya yang berhubungan dengan penelitian.

Data-data yang telah diperoleh dilakukan analisa data menggunakan metode deskriptif dengan pendekatan analisis data sekunder. Adapun teknis analisis danpenyajian hasil analisis data dilakukan dalam bentuk: Analisis perkembangan antar waktu (*time series*) hasil analisis ini disajikan dalam bentuk tabel dan atau grafik /diagram (diagram batang, diagram garis, atau gabungan diagram batang dan garis), Analisis posisi relatif hasil analisis disajikan dalam bentuk diagram/grafik batang ataugabungan diagram batang dan garis. Data yang telah diperoleh akan diolah dan diuji menggunakan model regresi Lasso dan Linier. Kedua model regresi tersebut akan dibandingkan dan dipilih model manakah yang lebih akurat dalam memprediksi tingkat kemiskinan di Indonesia.

HASIL DAN PEMBAHASAN

Data Acquisition/Selection

Tahap pertama pada penelitian ini adalah *selection. Selection* merupakan tahapan awal berupa pemilihan sampel dari berbagai sumber. Data yang digunakan pada penelitian ini adalah data kemiskinan penduduk pada tahun 2017 hingga 2021. Data diperoleh melalui dataset yang sudah disediakan oleh (Statistik, 2022).

Preprocessing

Data preprocessing merupakan langkah dalam proses *data mining* dan analisis data. Dalam proses ini, data mentah dipanggil dan disiapkan dalam bentuk yang dapat dipahami dan dianalisis oleh komputer. Hal ini diperlukan karena data mentah di dunia nyata, baik berupa teks, gambar, maupun video, masih berantakan. Oleh karena itu, sulit bagi komputer untuk memprosesnya.

Preprocessing data adalah langkah pertama dalam membuat model untuk *machine learning* dan kecerdasan buatan. Proses ini mengubah data menjadi bentuk yang lebih mudah dan efisien untuk diproses, sehingga *machine learning* dan pengembangan kecerdasan buatan memberikan hasil yang lebih akurat. Terdapat 4 langkah dalam melakukan *preprocessing data*, yaitu:

a. ***Data Cleaning*****:** Langkah ini merupakan tahapan awal dalam *preprocessing data*. Tahapan ini bertujuan untuk melakukan seleksi serta membuang data yang berpotensi mengurangi akurasi dari *machine learning* atau kecerdasan buatan. Beberapa permasalahan yang sering ditemui dalam tahap ini adalah *missing value, noisy data,* dan *inconsistent data*.

b. ***Data Integration*****:** *Data Integration* adalah suatu proses yang bertujuan untuk mengintegrasikan data yang berasal dari berbagai sumber menjadi satu data yang terpadu (*integrated data*). Proses ini meliputi pengumpulan, pemindahan, transformasi, dan penyatuan data yang berasal dari sumber-sumber yang berbeda. Namun dalam tahapan ini perlu dipastikan bahwa beberapa data yang digabungkan memiliki format yang sama. Apabila ditemukan perbedaan dalam format data yang digabungkan, maka ada beberapa cara yang bisa dilakukan yaitu memastikan data memiliki atribut dan format yang sama, menghapus atribut yang tidak diperlukan, serta mendeteksi nilai yang ambigu.

c. ***Data Transformation*****:** *Data Transformation* adalah proses mengubah struktur data dari satu format ke format lainnya. Transformasi data adalah proses pembersihan, pengubahan, dan pemodelan data agar dapat digunakan untuk tujuan yang berbeda. Proses ini bertujuan untuk mengubah data dari satu format menjadi format yang lainnya. Transformasi data memungkinkan data dari berbagai sumber dan format yang berbeda digabungkan, dimodelkan, dan dianalisis sebagai satu kesatuan.

d. ***Data Reduction*****:** Data Reduction adalah salah satu teknik untuk mengkompres data sehingga data dapat dengan mudah ditransmisikan dan disimpan. Teknik ini dapat digunakan untuk mengurangi ukuran file dan mengurangi ukuran data untuk meningkatkan kecepatan transfer data. Data reduction dapat dilakukan dengan menggunakan algoritma kompresi atau menggunakan teknik lain seperti de-duplikasi atau menghapus data yang tidak perlu. Teknik ini sangat berguna untuk mengurangi ukuran file yang besar dan meningkatkan kecepatan transfer data. Data reduction juga dapat digunakan untuk mengurangi waktu yang diperlukan untuk menyimpan data.

Transformation

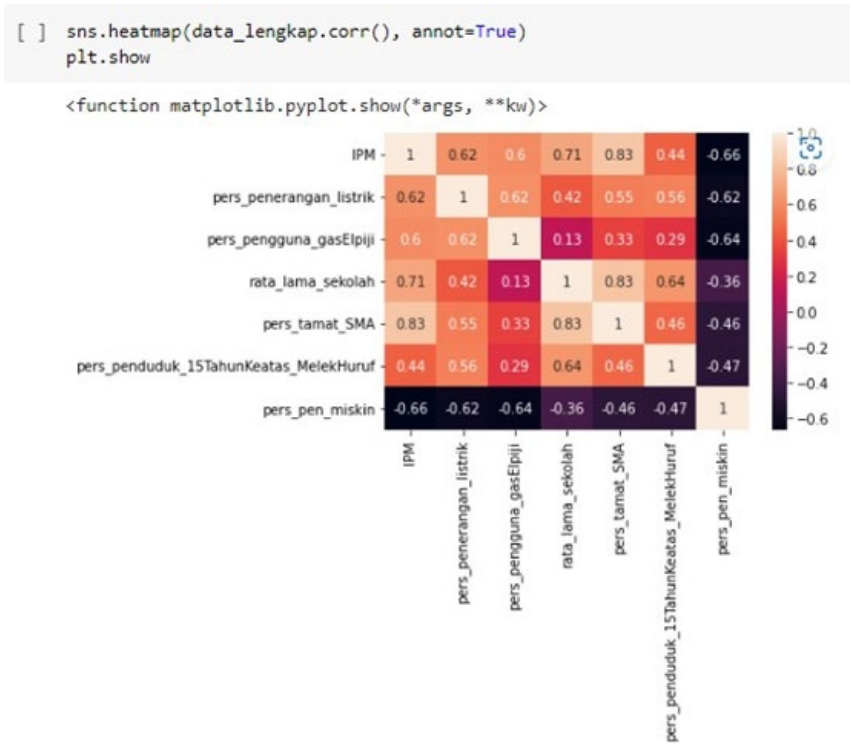
Transformasi data adalah proses pengubahan atau pengolahan data dari suatu bentuk ke bentuk lainnya. Transformasi data bertujuan untuk mengubah format data sehingga lebih mudah diproses oleh sistem komputer atau bisa juga untuk mengubah format data agar lebih sesuai dengan kebutuhan pengguna. Transformasi data dapat dilakukan dengan berbagai metode, seperti melakukan konversi bentuk data, penggabungan data, pemilihan data, pemotongan data, dan lain-lain. Transformasi data dapat dilakukan dengan menggunakan berbagai teknik, seperti regresi, klasifikasi, dan clustering. algoritma yang digunakan pada data transformation kali ini adalah regresi Lasso dan regresi linier. Regresi Lasso adalah sebuah teknik pembelajaran mesin yang digunakan untuk menentukan hubungan antara variabel yang saling terkait. Regresi linier adalah algoritma yang digunakan untuk memprediksi nilai suatu variabel tergantung pada nilai variabel lainnya.

Regresi linier mengasumsikan bahwa terdapat hubungan linier antara variabel tergantung dan variabel bebas, yaitu hubungan yang dapat dinyatakan dalam bentuk persamaan garis lurus. Kedua algoritma ini dapat digunakan untuk menganalisa dan memprediksi hubungan antara variabel-variabel dalam suatu dataset. Regresi Lasso menggunakan teknik penalisasi untuk mengurangi kemungkinan overfitting pada model. Overfitting terjadi ketika model terlalu disesuaikan dengan data training, sehingga tidak dapat menangani data baru dengan baik. Regresi Lasso memperkenalkan penalisasi pada model dengan mengurangi bobot variabel yang tidak signifikan secara bertahap, sehingga model menjadi lebih general dan dapat digunakan untuk data baru. Kedua algoritma regresi ini dapat digunakan untuk mengolah data dalam berbagai bidang,

seperti ekonomi, kesehatan, dan ilmu pengetahuan data lainnya. Regresi Lasso dan regresi linier dapat memberikan informasi yang bermanfaat tentang hubungan antara variabel-variabel dalam suatu dataset dan dapat digunakan untuk memprediksi nilai suatu variabel berdasarkan nilai variabel lainnya.

Data Mining

Data mining adalah suatu proses menggali informasi atau wawasan yang berguna dari suatu data yang besar, terstruktur atau tidak terstruktur. *Data mining* memanfaatkan teknik-teknik statistik dan matematika untuk menemukan pola atau korelasi dalam data yang dapat digunakan untuk meningkatkan pemahaman tentang data tersebut dan memprediksi perilaku atau kejadian di masa yang akan datang. Sebelum melakukan tahap selanjutnya, perlu dilakukan pemetaan menggunakan *heatmap* seperti yang ditunjukkan pada Gambar 2 sebagai berikut:



Gambar 2.
Heatmap untuk Mengetahui Nilai Korelasi Antar Variabel

Pada tahapan ini, selanjutnya dilakukan perbandingan 2 algoritma yaitu regresi Lasso dan regresi linier. Perbandingan dilakukan dengan cara mengolah dataset menggunakan algoritma di atas, di atas satu per-satu. sehingga dapat diketahui nilai MSE dan R^2 Score dari masing-masing masing algoritma. Regresi Lasso dan regresi linier adalah dua algoritma yang digunakan dalam data mining untuk memprediksi perilaku atau kejadian di masa yang akan datang berdasarkan pola atau korelasi dalam data yang telah dianalisis. Regresi Lasso menggunakan metode penalisasi L1 untuk mengurangi jumlah variabel yang digunakan dalam model regresi, sedangkan regresi linier menggunakan seluruh variabel yang tersedia dalam model. Kedua algoritma ini dapat digunakan untuk menghitung Mean Squared Error (MSE) dan R^2 Score dari data yang telah dianalisis, yang dapat digunakan untuk membandingkan kinerja kedua algoritma tersebut. Regresi linier mencari hubungan linier antara variabel terikat dan variabel tidak terikat dengan menggunakan persamaan garis yang dapat dituliskan sebagai:

$$\mathbf{Y = a + bX}$$

di mana Y adalah variabel terikat, X adalah variabel tidak terikat, dan a dan b adalah parameter yang ditentukan oleh model. Regresi Lasso adalah salah satu varian dari regresi linier

yang menggunakan regularisasi untuk mengurangi jumlah variabel yang digunakan dalam model dan menghindari *overfitting*. Regresi Lasso dapat dituliskan sebagai:

$$Y = a + b1X1 + b2X2 + ... + bnXn$$

di mana Y adalah variabel terikat, X1, X2, ..., Xn adalah variabel tidak terikat, dan a, b1, b2, ..., bn adalah parameter yang ditentukan oleh model.

Evaluation

Pada proses ini, dilakukan pengujian terhadap model yang telah dibangun sebelumnya menggunakan data uji atau data yang belum pernah digunakan sebelumnya dalam proses pembuatan model. Dengan menggunakan data uji tersebut, diperoleh performa dari masing-masing algoritma yang kemudian dibandingkan untuk menentukan metode mana yang lebih efektif dalam meramalkan tingkat kemiskinan. Pada tahap evaluasi, kedua metode tersebut akan diuji kemampuannya dalam memprediksi tingkat kemiskinan pada masing-masing provinsi di Indonesia.

Berdasarkan paparan di atas, didapatkan hasil dari pengolahan data menggunakan 2 metode algoritma yaitu Regresi Lasso dan Regresi linier, kemudian dibandingkan nilai MSE (*Mean Squared Error*) dan nilai R^2 pada masing-masing algoritma.

```
#MSE model Regresi Linier
print('Nilai MSE data training Regresi Linier = ', mean_squared_error(Y_train, Ypredtrain_reglin))
print('Nilai MSE data testing Regresi inier = ', mean_squared_error(Y_test, Ypredtest_reglin), '\n')

#MSE model Lasso
print('Nilai MSE data training Regresi Lasso = ', mean_squared_error(Y_train, Ypredtrain_lasso))
print('Nilai MsE data testing Regresi Lasso = ', mean_squared_error(Y_test, Ypredtest_lasso), '\n')
```

```
Nilai MSE data training Regresi Linier =  5.659936580399228
Nilai MSE data testing Regresi inier =  5.836056699216274

Nilai MSE data training Regresi Lasso =  5.728319684789296
Nilai MsE data testing Regresi Lasso =  6.302130948237999
```

Gambar 3.
Perbandingan Nilai MSE model Regresi Linier dan Regresi Lasso

Gambar 3 merupakan nilai MSE pada masing-masing model algoritma. Dari gambar di atas dapat diketahui bahwa Algoritma Regresi linier adalah algoritma yang baik, karena semakin kecil nilai MSE, maka semakin baik dalam memprediksi sesuatu. Regresi linier memiliki nilai data training 5.6 dan data testing 5.8.

```
[ ]  print(f'R^2 score Regresi Linier : {LinearReg.score(X, Y)}')
     print(f'R^2 score Regresi Lasso : {LassoReg.score(X, Y)}')
```

```
R^2 score Regresi Linier : 0.6348022045133301
R^2 score Regresi Lasso : 0.6262075398970801
```

Gambar 4.
Perbandingan Nilai R^2 pada Model Regresi Linier dan Regresi Lasso

Gambar 4 merupakan nilai R^2 pada masing-masing model algoritma. Hasil uji pada gambar di atas menunjukkan bahwa Algoritma Regresi linier merupakan metode terbaik karena apabila nilai R^2 semakin mendekati 1 maka semakin baik. Regresi linier memiliki nilai R^2 0.63.

KESIMPULAN DAN SARAN

Kesimpulan

Berdasarkan hasil penelitian dan pembahasan yang telah dilakukan dan dijabarkan di atas, dapat ditarik kesimpulan sebagai berikut: Diantara kedua algoritma, yaitu algoritma

Regresi Linier dan Regresi Lasso, yang memiliki tingkat akurasi lebih tinggi untuk memprediksi tingkat kemiskinan pada masing-masing provinsi di Indonesia adalah algoritma Regresi Linier karena memiliki nilai MSE yang lebih rendah serta memiliki nilai R^2 mendekati 1, dibandingkan algoritma regresi Lasso. Hasil analisis menunjukkan bahwa variabel yang memiliki pengaruh tertinggi terhadap tingkat kemiskinan pada provinsi di Indonesia yaitu pendidikan, serta Indeks Pembangunan Manusia (IPM).

Saran

Menggunakan model regresi atau model algoritma lainnya untuk memprediksi tingkat kemiskinan di Indonesia agar hasilnya dapat dibandingkan dan dipilih yang paling akurat. Pemerintah harus memperhatikan daerah-daerah pedesaan yang belum mengalami kemajuan baik secara ekonomi maupun sarana prasarana agar dapat memperbaiki tingkat kemiskinan di Indonesia.

DAFTAR PUSTAKA

Han, J., & Kamber, M. (2006). Data Mining: Concepts And Techniques, 2nd. *University Of Illinois At Urbana Champaign: Morgan Kaufmann*.

Permana, A. Y., & Arianti, F. (2012). Analisis Pengaruh Pdrb, Pengangguran, Pendidikan, Dan Kesehatan Terhadap Kemiskinan Di Jawa Tengah Tahun 2004-2009. *Diponegoro Journal Of Economics*, *1*(1), 25–32.

Pramudiono, I. (2007). *Pengantar Data Mining: Menambang Permata Pengetahuan Di Gunung Data*.

Santosa, B. (2007). Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis. *Yogyakarta: Graha Ilmu*, *978*(979), 756.

Setiadi, E. M., & Kolip, U. (2011). Pengantar Sosiologi (Jakarta. *Kencana Prenada Media Group*.

Statistik, B. P. (2011). Sumatera Barat Dalam Angka. *Badan Perencanaan Dan Pembangunan Daerah Sumbar. Padang*.

Statistik, B. P. (2018). Indeks Pembangunan Manusia (Ipm) Tahun 2018. *Berita Resmi Statistik, Available At: Https://Doi. Org/4102002*.

Statistik, B. P. (2022). *Bps Provinsi Dki Jakarta*.

Suharto, E. (2009). *Membangun Masyarakat Memberdayakan Rakyat*.

Suryawati, C. (2005). Memahami Kemiskinan Secara Multidimensional. *Jurnal Manajemen Pelayanan Kesehatan*, *8*(03).

Todaro, M. P., & Smith, S. C. (2004). Pembangunan Ekonomi Di Dunia Ketiga Edisi Kedelapan. *Jakarta: Penerbit Erlangga*.

Witten, I. H., & Frank, E. (2002). Data Mining: Practical Machine Learning Tools And Techniques With Java Implementations. *Acm Sigmod Record*, *31*(1), 76–77.