

INTISARI

PERBANDINGAN ALGORITMA CGD LASSO DENGAN LARS

LASSO UNTUK MENGANALISIS DATA BERDIMENSI

TINGGI

Oleh

Vincentius Ryo siswoyo

18/427714/PA/18674

Data sudah menjadi dasar dari seluruh operasi statistik. Semua analisis pasti membutuhkan data. Pada analisis tersebut ada banyak asumsi-asumsi berhubungan dengan data yang harus dipenuhi. Contoh analisis tersebut adalah analisis regresi. Salah satu asumsi yang harus dipenuhi dalam analisis regresi adalah asumsi non-multikolinearitas, yaitu asumsi bahwa setiap variabel independen pada data tidak mempengaruhi variabel independen lainnya. Namun pada praktiknya tidak semua data dapat memenuhi asumsi tersebut. Contoh data yang memiliki masalah tersebut adalah data yang digunakan pada penelitian ini adalah data covid-19. Data berdimensi tinggi ($p > n$) tersebut memiliki masalah multikolinearitas sehingga tidak dapat dianalisis menggunakan metode regresi biasa. Maka untuk itu perlu dilakukan metode alternatif yang dapat digunakan pada data yang tidak memenuhi asumsi tersebut. Salah satu metode alternatif tersebut adalah analisis LASSO. Analisis LASSO adalah metode regresi alternatif yang dapat digunakan untuk mengatasi masalah multikolinearitas. Dalam metode LASSO ada berbagai algoritma yang dapat digunakan sesuai dengan situasi dan kondisi. Algoritma yang digunakan dalam penelitian ini adalah algoritma *gradient descent* dan LARS. Tujuan dari penelitian ini adalah untuk menentukan algoritma LASSO yang paling efektif dalam menganalisis data covid-19. Untuk menentukan metode yang terbaik maka dilihat dari nilai MSE dan MAE terkecil, dan nilai R^2 terbesar. Data covid-19 yang digunakan adalah data pada tanggal 1 februari 2022 sampai tanggal 15 maret 2022. Berdasarkan metrik evaluasi yang diperoleh dapat dilihat bahwa nilai metrik evaluasi algoritma LARS lebih baik dari algoritma lainnya. Hal tersebut

mengindikasikan bahwa metode LARS LASSO adalah metode yang paling efektif untuk menganalisis data covid-19. Dengan menggunakan metode LARS LASSO, didapatkan hasil bahwa faktor faktor yang mempengaruhi nilai *reproduction rate* adalah *new cases*, *new deaths smoothed*, *total deaths per million*, *total tests per thousand*, *new tests*, *new tests smoothed*, *positive rate*, *total vaccinations*, dan *new vaccinations smoothed*.

Kata Kunci : *Covid-19*, *LASSO*, *LARS LASSO*, *CGD LASSO*

ABSTRACT

COMPARISON OF CGD LASSO WITH LARS LASSO ALGORITHM TO ANALYZE HIGH DIMENSIONAL DATA

By

Vincentius Ryo Siswoyo

18/427714/Pa/18674

Data is the basis of all statistical operations. All analysis needs data. In statistical analysis there are many assumptions related to the data that must be met. An example of such analysis is regression analysis. One of the assumptions that must be met in regression analysis is the assumption of non-multicollinearity, namely the assumption that each independent variable in the data does not affect the other independent variables. However, in practice not all data can meet these assumptions. An example of data that has this problem is the data used in this study, the data on COVID-19. This high-dimensional data ($p > n$) has a multicollinearity problem so that it cannot be analyzed using the usual regression method. Therefore, it is necessary to use alternative methods that can be used on data that do not meet these assumptions. One such alternative method is LASSO analysis. LASSO analysis is an alternative regression method that can be used to solve the problem of multicollinearity. In the LASSO method there are various algorithms that can be used according to the situation. The algorithm used in this study is the gradient descent algorithm and LARS. The purpose of this study was to determine the most effective LASSO algorithm in analyzing COVID-19 data. To determine the best method, it is seen from the smallest MSE and MAE values, and the largest R^2 value. The covid-19 data used is data from February 1, 2022 to March 15, 2022. Based on the evaluation metric obtained, it can be seen that the evaluation metric value of the LARS algorithm is better than other algorithms. This indicates that the LARS LASSO method is the most effective method for analyzing COVID-19 data. By using the LARS LASSO method, it is found that the factors that affect the reproduction rate are new cases, new deaths smoothed, total deaths per million, total

tests per thousand, new tests, new tests smoothed, positive rate, total vaccinations,
and new vaccinations. smoothed.

Keyword: *Covid-19, LASSO, CGD LASSO, LARS LASSO*