INTISARI

ALGORITMA REGRESI *ROBUST* DAN *DECISION TREE* UNTUK PREDIKSI PADA DATASET DENGAN *OUTLIER*

Oleh

Salsabila Basalamah

21/485763/PPA/06226

Dalam dataset penelitian sering kali terjadi masalah seperti adanya data *outlier*. *Outlier* merupakan kumpulan objek dalam dataset yang berbeda, dari kumpulan objek-objek lain dalam suatu dataset. Perbedaan ini, dapat mengakibatkan hasil analisis menjadi bias. Sehingga, adanya *outlier* dalam dataset sering kali dihilangkan dari kumpulan data. Menghilangkan *outlier* dalam kumpulan dataset, memiliki kemungkinan resiko hilangnya informasi penting pada kumpulan data. Sebab *outlier* pada kondisi dan kasus tertentu bisa memiliki pengaruh yang cukup kuat dalam kumpulan datanya.

Hasil analisis yang bias akibat adanya *outlier* ini, dikarenakan algoritma yang membangun metode untuk melakukan prediksi tersebut tidak cukup kuat dalam mengatasi permasalahan *outlier* dan dalam beberapa metode mensyaratkan tidak adanya data *outlier*. Sehingga, dibutuhkan algoritma yang mampu melakukan prediksi untuk data *outlier* dan tidak adanya tindakan menghilangkan data *outlier*. Salah satu algoritma yang tepat untuk mengatasi masalah *outlier* ini adalah algoritma Regresi *Robust*. Regresi *Robust* dimasukkan ke dalam salah satu analisis *Machine Learning* yaitu *Decision Tree* Regresi. Dimana menggunakan Regresi *Robust* estimasi M Huber dan M Tukey Bisquare untuk menghasilkan prediksi pada dataset dengan *outlier* dan mencegah adanya bias pada hasil analisis, tanpa melakukan penghilangan data *outlier*.

Penelitian ini menggunakan lima dataset regresi yang bersumber dari *University of California Irvine* (UCI). Tiga dari lima dataset memperoleh hasil metode *Decision Tree* Regresi dengan M Huber memberikan prediksi yang lebih baik, pada dataset *Concrete*, dataset *Superconductivity* dan dataset *Airfoil* memiliki *Mean Absolute Error* (MAE) secara berturut-turut yaitu 3,963, 9,140, dan 1,644. Sedangkan, dua dataset lain lebih baik menggunakan estimasi OLS dan estimasi M Tukey Bisquare.

Kata Kunci: *Decision Tree* Regresi, *Estimasi M Huber, Estimasi M Tukey Bisquare, Outlier, Regresi Robust.*

ABSTRACT

ROBUST AND DECISION TREE REGRESSION ALGORITHMS FOR PREDICTION ON DATASETS WITH OUTLIER

Salsabila Basalamah

21/485763/PPA/06226

Problems often occur in research datasets, such as data with outliers. An outlier is a case that describes the characteristics of a difference. This difference can cause the results of the analysis to be biased. Thus, removing outliers in the data set is often done. Removing outliers in a data set has the risk of losing important information in the data set because outliers in certain conditions and cases can strongly influence the data set.

The analysis results are biased due to these outliers because the algorithm that builds the method for making these predictions is not strong enough to overcome the problem of outliers. In some models, is no sensitivity to data with the outlier. So, we need an algorithm that can make predictions for outlier data, and there is no action to remove outlier data. One of the appropriate algorithms to overcome this outlier problem is the Robust Regression algorithm. Robust regression is included in one of the analyses of Machine Learning, namely Decision Tree Regression. It employs M Huber and M Tukey Bisquare Robust Regression techniques for generating predictions on datasets containing outliers without eliminating the outlier data. This approach helps prevent bias in the analysis results.

This study utilizes five regression datasets from the University of California Irvine (UCI). Among the five datasets, the Regression Decision Tree method with M Huber provided better predictions for three of them. Specifically, in the Concrete datasets, Superconductivity datasets, and Airfoil datasets, the Mean Absolute Errors (MAE) were 3.963, 9.140, and 1.644, respectively. Conversely, the remaining two datasets exhibited better results using OLS estimates and M Tukey Bisquare estimates.

Keyword: *Decision Tree Regression, Huber's M-estimator, Outliers, Robust Regression, Tukey Bisquare's M-estimator*.