

ABSTRAK

Dalam era digital saat ini, platform seperti *youtube* menghasilkan volume data yang sangat besar, menciptakan tantangan signifikan dalam mengelompokkan data secara efektif untuk mengidentifikasi pola dan informasi berharga. Penelitian ini bertujuan untuk menentukan metode *clustering* yang optimal antara *K-Means* dan *DBSCAN clustering* dalam pengelompokan data *statistic global youtube* berdasarkan karakteristik seperti jumlah '*subscribers*', '*videoviews*', '*uploads*', '*video_views_for_the_last_30_days*', '*lowest_monthly_earnings*', '*highest_monthly_earnings*', '*lowest_yearly_earnings*', '*highest_yearly_earnings*', '*subscribers_for_last_30_days*', '*Population*', '*Urban_Population*', serta membandingkan kinerja kedua metode tersebut. Algoritma *K-Means* digunakan dengan parameter $K=3$, yang menghasilkan *silhouette coefficient* sebesar 0.70, menunjukkan tiga *cluster* yang terbentuk dengan waktu eksekusi 0.74 detik. Sementara itu, algoritma *Density-Based Spatial Clustering of Applications with Noise (DBSCAN)* digunakan dengan parameter *Eps* 0.8 dan *MinPts* 7, menghasilkan *silhouette coefficient* sebesar 0.72 dengan waktu eksekusi 0.10 detik. *DBSCAN* berhasil mengidentifikasi dua *cluster* dan *outlier* yang dikelompokkan ke dalam *cluster -1*. Hasil penelitian menunjukkan bahwa *DBSCAN* lebih unggul dibandingkan *K-Means* dalam konteks *dataset* yang diuji, karena menghasilkan *cluster* yang lebih padat dan terpisah dengan baik serta mampu menangani *outlier*. Berdasarkan penelitian ini, *DBSCAN* dianggap sebagai metode *clustering* yang lebih akurat dan efektif untuk *dataset* berdasarkan *silhouette coefficient* dan waktu eksekusi. Penelitian ini menyarankan penambahan antarmuka pengguna grafis untuk mempermudah proses penginputan parameter dan eksplorasi metode imputasi data selain *mean* serta menggunakan metode *k-distance graph* untuk penentuan parameter *eps* kepada penelitian selanjutnya.

Kata Kunci: *K-Means, Density Based DBSCAN, Youtube, Clustering*

ABSTRACT

In the current digital era, platforms like YouTube generate vast amounts of data, posing significant challenges in effectively clustering this data to identify valuable patterns and information. This research aims to determine the optimal clustering method between K-Means and DBSCAN for grouping *statistic global youtube* data based on characteristics such as 'subscribers', 'videoviews', 'uploads', 'video_views_for_the_last_30_days', 'lowest_monthly_earnings', 'highest_monthly_earnings', 'lowest_yearly_earnings', 'highest_yearly_earnings', 'subscribers_for_last_30_days', 'Population', and 'Urban_Population', and to compare the performance of these two methods. The K-Means algorithm was used with a parameter of $K=3$, yielding a silhouette coefficient of 0.70 and forming three clusters with an execution time of 0.74 seconds. Meanwhile, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used with parameters $Eps=0.8$ and $MinPts=7$, resulting in a silhouette coefficient of 0.72 and an execution time of 0.10 seconds. DBSCAN successfully identified two clusters and outliers grouped into cluster -1. The results indicate that DBSCAN outperforms K-Means in the context of the tested dataset, as it produces more compact and well-separated clusters and effectively handles outliers. Based on this research, DBSCAN is considered a more accurate and effective clustering method for datasets, based on the silhouette coefficient and execution time. This research recommends adding a graphical user interface to facilitate parameter input and exploring data imputation methods other than the mean. Additionally, it suggests using the k-distance graph method for determining the Eps parameter in future studies.

Key Word: *K-Means, Density Based, Youtube, Clustering*