Abstrak

Preprocessing di dalam *data mining* adalah salah satu faktor penting dalam menyiapkan data sehingga menghasilkan informasi yang efisien dan berkualitas. Pada *unsupervised learning* atau *clustering*, pemrosesan data berdimensi tinggi akan membutuhkan biaya dan waktu komputasi yang besar. Proses *clustering* pun dapat bekerja lebih baik pada data yang berdimensi sedikit.

Teknik *preprocessing* yang dibahas pada tugas akhir ini adalah *Principal Component Analysis* (PCA) dimana data set yang dimensinya besar diringkas menjadi data set dengan dimensi baru yang jumlahnya lebih sedikit. Dimensi yang baru disebut *principal component* (PC). PC dibentuk dari kombinasi linier dari dimensi asli sehingga data tidak akan kehilangan karakteristik aslinya.

Hasil pengujian sistem menghasilkan data *colon tumor* dengan 2000 dimensi dapat diringkas menjadi 60 PC dan data set DLBCL dengan 4026 dimensi dapat diringkas menjadi 46 PC. Pada data set *colon tumor* dan DLBCL, data 1, 2, atau 3 PC dapat memberikan performansi hasil *K-Means Clustering* yang lebih baik daripada data asli. Untuk metode *Two Step Clustering* pada data set *colon tumor* diperoleh performansi PCA yang kurang efektif sedangkan pada data set DLBCL diperoleh performansi PCA yang baik pada data 1 atau 3 PC.

Kata kunci: *data mining, preprocessing,* PCA, *clustering,* dimensi tinggi