Clustering Varian Sekuens Protein Sars-Cov-2 Menggunakan Algoritma Spectral = Clustering of Variants of the SARS-CoV-2 Protein Sequence Using Spectral Algorithm

Azkal Azkiya, author

Abstrak

Coronavirus disease (COVID-19) adalah penyakit pernapasan menular yang disebabkan oleh jenis coronavirus baru. Penyakit ini sebelumnya disebut dengan 2019-nCoV atau 2019 novel coronavirus. Virus penyebab COVID-19 ini adalah SARS-CoV-2. Terdapat varian SARS-CoV-2 lain yang memiliki potensi berdampak besar bagi kesehatan masyarakat seperti Lambda dan Mu. Ada pula kelompok varian SARS-CoV-2 under monitoring yang belum diketahui dampak dan bentuk penyebarannya di tingkat masyarakat. Kappa, Iota, dan Epsilon merupakan beberapa contoh varian yang termasuk ke dalam kelompok tersebut. World Health Organization (WHO) terus melakukan pengawasan kemunculan varian SARS-CoV-2 yang baru. Varian SARS-CoV-2 yang telah diketahui penularan dan dampaknya cukup signifikan pada masyarakat hingga saat ini adalah Alpha, Beta, Delta, Gamma, dan Omicron. Penelitian ini menggunakan data dari kelima varian SARS-CoV-2 tersebut. Penelitian ini mengimplementasikan program unsupervised dari machine learning yaitu simulasi proses clustering untuk mengelompokkan varian SARS-CoV-2. Dilakukan ekstraksi fitur terhadap data sekuens protein SARS-CoV-2 menggunakan package discere dalam bahasa pemrograman Python. Melalui proses ekstraksi fitur dihasilkan 27 fitur data sekuens protein SARS-CoV-2 yang siap digunakan. Elbow method kemudian diimplementasikan terhadap data untuk mengetahui jumlah pembentukan cluster yang optimal untuk digunakan pada clustering. Berdasarkan elbow method didapatkan jumlah cluster optimal untuk simulasi clustering sebanyak dan dilakukan juga simulasi dengan untuk memberi kesempatan kepada seluruh varian untuk membentuk clusternya sendiri. Metode clustering yang digunakan pada penelitian ini adalah spectral clustering. Cluster yang dihasilkan kemudian dievaluasi menggunakan metrik evaluasi silhouette score serta melihat runtime pada setiap simulasi yang dilakukan. Hasil silhouette score untuk simulasi dengan bernilai 0,614 dan untuk simulasi dengan yang bernilai 0,631. Durasi rata-rata runtime mencatat bahwa simulasi dengan dengan 6,566 detik lebih baik dibanding simulasi dengan dengan 7,529 detik. Berdasarkan hasil tersebut, spectral clustering dapat dilakukan terhadap varian SARS-CoV-2 dengan pemilihan jumlah cluster menggunakan elbow method.

.....

Coronavirus disease (COVID-19) is an infectious respiratory disease caused by a new type of coronavirus. This disease was previously called 2019-nCoV or 2019 novel coronavirus. The virus that causes COVID-19 is the SARS-CoV-2. There are several variants of SARS-CoV-2 that have the potential to have a major impact on public health, such as Lambda and Mu. There is also a group of variants of SARS-CoV-2 under monitoring whose impact and form of spread are unknown at the community level.

Kappa, Iota, and Epsilon are some examples of variants that belong to this group. The World Health Organization (WHO) continues to monitor the emergence of a new variant of SARS-CoV-2. The variants of SARS-CoV-2 that are known to transmit and have a significant impact on society so far are Alpha, Beta, Delta, Gamma and Omicron. This study uses data from that five variants of SARS-CoV-2. This study implements an unsupervised program from machine learning, which is a simulation of the clustering process to group variants of SARS-CoV-2 . Feature extraction was carried out on the SARS-CoV-2 protein sequence data using discere package in the Python programming language. Through the feature extraction process, 27 features of the SARS-CoV-2 protein sequence data were produced which were ready for use. The elbow method is then implemented on the data to find out the optimal number of cluster formations for use in clustering. Based on the elbow method, the optimal number of clusters for the clustering simulation is and a simulation with is also carried out to provide an opportunity for all variants to form their own clusters. The clustering method used in this study is spectral clustering. The resulting clusters are then evaluated using the silhouette score evaluation metric and looking at the runtime in each simulation that is performed. The results of the silhouette score for the simulation with is worth 0.614 and for the simulation with it is worth 0.631. The average duration of the runtime noted that the simulation with with 6.566 seconds was better than the simulation with with 7.529 seconds. Based on these results, spectral clustering can be carried out on the SARS-CoV-2 variant by selecting the number of clusters using the elbow method.