Klasifikasi Sekuens Protein Coronavirus Penyebab COVID-19 Menggunakan Metode LightGBM dengan Seleksi Fitur Elastic Net = Coronavirus Protein Sequence Classification Causes of COVID-19 Using the LightGBM Method with Elastic Net Feature Selection

Febiola Damayanti, author

Abstrak

Pandemi COVID-19 (coronavirus disease 2019) membuat para peneliti di seluruh dunia bekerja untuk memahaminya dengan menerapkan pendekatan machine learning. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) merupakan penyebab dari COVID-19. Penelitian ini membahas klasifikaasi sekuens protein SARS-CoV-2 menggunakan metode LightGBM dan Elastic Net. Metode LightGBM merupakan metode gradient boosting yang cepat dan memiliki high-performance berbasis decision tree untuk melakukan prediksi. Total data sekuens protein yang digunakan adalah 2000 data yang diambil dari situs Uniprot. Uniprot merupakan salah satu situs yang digunakan terkait bioinformatika atau sumber daya sekuens protein dan informasi fungsional yang memiliki kualitas tinggi, komprehensif dan dapat diakses secara bebas. Data tersebut memiliki perincian yaitu 1000 data sekuens protein SARS-CoV-2 dan 1000 data sekuens protein bukan SARS-CoV-2. Python package Discere digunakan untuk mengekstraksi 27 fitur sekuens protein. Selanjutnya, Elastic Net digunakan untuk memilih fitur-fitur yang optimal dan terpilih sebanyak 10 fitur. Terakhir, LightGBM digunakan sebagai metode klasifikasi sekuens protein SARS-CoV-2. Hasil evaluasi performa LightGBM diukur dari akurasi, sensitivitas, dan spesifisitas. Nilai rata-rata akurasi diperoleh 98,87%, nilai rata-rata sensitivitas diperoleh 99,02%, dan nilai rata-rata spesifisitas diperoleh 98,82%

.....The COVID-19 (coronavirus disease 2019) pandemic has researchers around the world working to understand it by applying a machine-learning approach. Secere acute respiratory syndrome coronavirus 2 (SARS-Cov-2) is the cause of COVID-19. This research discusses the classification of SARS-Cov-2 protein sequences using the LightGBM and Elastic Net methods. The LightGBM method is a gradient-boosting method that fast and has a high-performance decision tree based for making predictions. The total protein sequence data used is 2000 data taken from UniProt site. UniProt is one of the sites used for bioinformatics or protein sequence resources and functional information which is of high quality, comprehensive and freely accesible. The data has details, namely 1000 protein sequence data for SARS-CoV-2 and 1000 protein sequnce data for non-SARS-CoV-2. Python package Dsiscere is used to extraxt 27 protein sequence features. Futhermore, Elastic Net is used to select optimal features and 10 features are selected. While LightGBM is used as a classification method for SARS-Cov-2 protein sequences. The results of the LightGBM performance evaluation are measured by accuracy, sensitivity, and specificity. The average value for accuracy was 98,87%, the average value for sensitivity was 99,02%, and average value for specificity was 98,82%.