ABSTRAK

Laporan yang berjudul "Analisis Akurasi Data Protein Virus SARS-CoV-2 dengan menggunakan Metode *Data mining*" ini bertujuan untuk menganalisis hasil penelitian *data mining* pada kumpulan data protein SARS-CoV-2 antara algoritma Naïve Bayes dan J48 dengan pelatihan *10-fold cross-validation*. Proses penelitian ini dimulai dengan pengumpulan data yang disediakan oleh *website* NCBI berupa protein virus SARS-CoV-2. Selanjutnya pada tahap praproses data, kumpulan data tersebut dibersihkan, data yang akan digunakan dari jumlah kumpulan data 1.149.216 menjadi 517.834. Dengan aplikasi WEKA, *data mining* algoritma Naïve Bayes dan J48 akan diterapkan pada kumpulan data yang akan mengahsilkan model klasifikasi. Selanjutnya, dengan 500 data uji dari *website* NCBI, model klasifikasi akan memprediksikan kelas protein pada data uji. Pada tahap analisis hasil data, akan dibandingkan kinerja masing-masing algoritma dari hasil detail akurasi juga waktu pembuatan model dan pengujian data. Dapat disimpulkan dengan menggunakan algoritma J48 lebih unggul akurasinya dibanding Naïve Bayes dalam melakukan klasifikasi pada kumpulan data protein SARS-CoV-2. Laporan ini dibuat sebagai pengenalan awal bioinformatika dan proses melakukan *data mining* pada kumpulan data bioinformatika.

Kata kunci: Bioinformatika, *Data mining*, J48, Naïve Bayes, SARS-CoV-2.

ABSTRACT

The report entitled "Analisis Akurasi Data Protein Virus SARS-CoV-2 dengan menggunakan Metode Data mining" aims to analyze the result of data mining research on the SARS-CoV-2 protein data set between the Naïve Bayes and J48 algorithms with 10-fold cross-validation training. The research process begins with the collection of data provided by the NCBI website about proteins SARS-CoV-2 virus. At the preprocessing stage, the data set is cleaned, from total data set of 1,149,216 to 517,834. With the WEKA application, the Naïve Bayes and J48 data mining algorithms will be applied to a data set that will produce a classification model. With 500 test data from the NCBI website, the classification model will predict the protein class. At the data analysis stage, the performance of each algorithm will be compared from the detailed results of the accuracy as well as the time for making the model and testing the data. From this research, it can be concluded that using J48 algorithm is more accurate than Naïve Bayes in classifying protein SARS-CoV-2 virus data set. This report was created as an introduction to bioinformatics and the process of data mining on bioinformatics data sets.

Keywords: Bioinformatics, Data mining, J48, Naïve Bayes, SARS-CoV-2.

DAFTAR ISI

Universitas Kristen Maranatha

Universitas Kristen Maranatha

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR SINGKATAN

ARFF	*Attribute Relation File Format*
CART	*Classification and Regression Trees*
CBIIT	*Center for Biomedical Informatics & Information Technology*
CSV	*Comma-Separated Value*
DDBJ	*DNA Database of Japan*
DNA	*Deoxyribonucleic Acid*
EMBL	*European Molecular Biology Laboratory*
Envelope protein	*The CoV E protein*
ID3	*Iterative Dichotomiser 3*
KDD	*Knowledge Discovery in Database*
KNN	*K-Nearest Neighbour*
Membrane glycoproteins	*membrane proteins*
NCBI	*National Center for Biotechnology Information*
Nucleocapsid phosphoprotein	*The multifunctional nucleocapsid protein*
ORF10 protein	*Open Reading Frame 10 protein*
ORF1a polyprotein	*Open Reading Frame 1a polyprotein*
ORF1ab polyprotein	*Open Reading Frame 1ab polyprotein*
ORF3a protein	*Open Reading Frame 3a protein*
ORF6 protein	*Open reading Frame 6 protein*
ORF7b protein	*Open Reading Frame 7 protein*
ORF8 protein	*Open Reading Frame 8 protein*
ORF10 protein	*Open Reading Frame 10 protein*

DAFTAR ISTILAH

Bioinformatika	berupa konsep dari biologi yang memandang segi molekul dan menerapkan teknik informatika untuk memahami informasi yang berkaitan dengan molekul dalam skala besar.
Data mining	merupakan metode untuk menggali suatu pola dari kumpulan data yang besar dan kompleks dengan menggunakan algoritma.
DNA	molekul kompleks yang berisi informasi yang diperlukan untuk membangun dan memelihara organisme.
J48	singkatan dari weka.classifiers.trees.j48 yang menggunakan algoritma Quinlan yaitu C4.5.
K-fold Cross-Validation	salah satu pelatihan klasifikasi di WEKA.
Naïve Bayes	klasifikasi yang menekankan kepada bidang statistika yaitu probabilitas.
RNA	molekul yang mempunyai peran dalam pengkodean, pengurai, regulasi dan ekspresi gen.