

ABSTRACT

FEATURE SELECTION IN C4.5 ALGORITHM AND RANDOM FOREST ALGORITHM IN DIAGNOSING CHRONIC KIDNEY DISEASE

By:
NAWA ISMAIL
2221210051

Chronic Kidney Disease (CKD) is characterized by decreased kidney function that causes the kidneys to be unable to remove toxins and waste products from the blood. The decreased kidney function is because the presence of protein in the urine and a decrease in the glomerular filtration rate. Chronic kidney disease contributes greatly to the mortality rate from year to year. High mortality rates caused by chronic kidney disease must be prevented immediately. One way to prevent this disease is to diagnose it with accuracy and precision. One of the methods that can be used for diagnosis it are the C4.5 algorithm and Random Forest algorithm. To increase the accuracy value in both algorithms, statistical-based feature selection is used. Based on feature selection and the concept of person correlation, the tool used in this research is Google collaboratory with Python language, through the stages of data pre-processing, feature selection, modelling the C4.5 algorithm, and the Random Forest algorithm. While for evaluation is done by dividing the data into training data and test data with a presentation of 80% for training data and 20% for test data. Evaluation results after feature selection increased the accuracy value by 1.25% from 96.25% to 97.50% in the C4.5 algorithm. In contrast, the Random Forest algorithm increased the accuracy value by 1.25% from 98.75% to 100%.

Keywords: Chronic Kidney Disease, C4.5 Algorithm, Random Forest, Feature Selection

INTISARI

Seleksi Fitur Pada Algoritme C4.5 dan Algoritme *Random Forest* dalam Mendiagnosis Penyakit Ginjal Kronis

Oleh:

Nawa Ismail
2221210051

Penyakit Ginjal Kronik (PGK) merupakan keadaan terjadinya penurunan fungsi ginjal yang menyebabkan ginjal tidak dapat membuang racun dan produk sisa dari darah, yang ditandai adanya protein dalam urin serta penurunan laju filtrasi glomerulus. Penyakit ginjal kronis berkontribusi besar pada angka kematian dari tahun ke tahun. Tingginya kematian yang disebabkan oleh ginjal kronik harus segera dilakukan pencegahan, salah satunya dengan melakukan diagnosis penyakit ginjal kronis adalah secara tepat dan akurat. Salah satu metode yang dapat digunakan untuk diagnosis adalah algoritme C4.5. dan algoritme *Random Forest*. Untuk meningkatkan nilai akurasi pada kedua algoritma tersebut digunakan seleksi fitur berbasis statistik dengan konsep korelasi person. *Tool* yang digunakan pada penelitian ini adalah google colabatory dengan bahasa pyton, melalui tahapan pre-processing data, seleksi fitur, pemodelan algoritma C4.5 dan algoritme *Random Forest*. Sementara untuk evaluasi dilakukan dengan membagi data menjadi data latih dan data uji dengan presentasi 80% untuk data latih dan 20% untuk data uji. Hasil evaluasi setelah dilakukan seleksi fitur mengalami peningkatan nilai akurasi sebesar 1,25% dari 96,25% meningkat 97,50% pada algoritme C4.5 sedangkan pada algoritme *Random Forest* sebesar sebesar 1,25% dari 98,75% meningkat 100%.

Kata Kunci: Penyakit Ginjal Kronik, Algoritme C4.5, *Random Forest*, Seleksi Fitur