KLASIFIKASI MULTI-LABEL MENGGUNAKAN METODE *MULTI-LABEL K-NEAREST NEIGHBOR* (ML-KNN) PADA PENYAKIT KANKER SERVIKS

ERISA RIZKYANI

ABSTRAK

Berdasarkan data statistik GLOBOCAN 2020, kanker serviks menempati urutan ke-8 penyakit kanker paling banyak diderita perempuan di seluruh dunia. *Multi-Label K-Nearest Neighbor* (ML-KNN) termasuk dalam salah satu *adaptive algorithm* yang digunakan untuk menyelesaikan kasus klasifikasi multi-label. *Dataset* yang digunakan pada penelitian ini diperoleh dari *website UCI Machine Learning*. Pada *dataset* tersebut akan dilakukan pra-proses data dengan menghilangkan *missing value*, cek *duplicate data*, cek tipe data, dan *resample* data dengan melakukan *oversampling* pada label *Biopsy* karena data yang tidak seimbang. Setelah itu data dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Data latih dicari kedekatannya dengan nilai k yang sudah ditentukan yaitu K=1, K=3, K=5, K=7, dan K=9. Hasil evaluasi diperoleh performa terbaik untuk klasifikasi ML-KNN yaitu saat nilai K=5 yang memperoleh nilai *hamming loss* sebesar 3,59%, akurasi sebesar 93%, *precision weighted* sebesar 93%, *recall weighted* sebesar 96%, dan *f1-score weighted* sebesar 94%.

Kata kunci: Klasifikasi, Kanker Serviks, *Multi-Label K-Nearest Neighbor* (ML-KNN), *Oversampling*.

MULTI-LABEL CLASSIFICATION USING THE MULTI-LABEL K-NEAREST NEIGHBOR (ML-KNN) ALGORITHM ON CERVIC CANCER

ERISA RIZKYANI

ABSTRACT

Based on GLOBOCAN 2020 statistical data, cervical cancer is the 8th most common cancer in women worldwide. Multi-Label K-Nearest Neighbor (ML-KNN) is one of the adaptive algorithms used to solve multi-label classification cases. The dataset used in this study was obtained from the UCI Machine Learning website. The dataset will be preprocessed by eliminating missing values, checking for duplicate data, checking data types, and resampling data by oversampling the Biopsy label due to unbalanced data. After that the data is divided into training data and test data with a ratio of 80:20. The training data is searched for its proximity to the predetermined k value, namely K=1, K=3, K=5, K=7, and K=9. The evaluation results obtained the best performance for the ML-KNN classification, namely when the value of K = 5 which obtained a hamming loss value of 3.59%, accuracy of 93%, precision weighted of 93%, recall weighted of 96%, and f1-score weighted of 94%.

***Keywords**: Classification, Cervical Cancer, Multi-Label K-Nearest Neighbor (ML-KNN), Oversampling.*