

ABSTRAK

Algoritma C4.5 merupakan metode klasifikasi dan prediksi dalam teknik penambangan data yang digunakan untuk mengidentifikasi pola tersembunyi dalam data untuk memperoleh informasi atau pengetahuan yang berguna dari suatu data yang berjumlah besar. Pada penelitian ini algoritma C4.5 diterapkan untuk mendeteksi penyakit kanker serviks dengan mengklasifikasi hasil biopsi pasien berdasarkan atribut yang mempengaruhi risiko kanker serviks dari *dataset* sebuah Rumah Sakit di Caracas, Venezuela. Data tersebut berjumlah 858 data dalam kondisi *imbalanced data* karena terdapat kelas yang mendominasi dari kelas lainnya. Masalah *imbalanced data* dapat menyebabkan hasil dari klasifikasi tidak akurat. Untuk mengatasi masalah tersebut, peneliti menerapkan teknik *undersampling* terhadap data mayoritas agar data yang dimiliki menjadi seimbang dan memperoleh hasil yang akurat.

Peneliti melakukan pengujian dengan jumlah 165 data. Pada pengujian dengan perbandingan data *training* dan data *testing* secara *descending* sebesar 80:20 menghasilkan tingkat akurasi tertinggi sebesar 75,76%. Berdasarkan hasil perhitungan dapat disimpulkan bahwa algoritma C4.5 dapat diterapkan dan menemukan aturan klasifikasi yang menarik dari pohon keputusan yang terbentuk.

Kata kunci : algoritma C4.5, *undersampling*, kanker serviks

ABSTRACT

C4.5 algorithm is a method of classification and prediction in data mining techniques that used to identify hidden patterns in data to obtain information or knowledge from a large amount of data. In this research, the C4.5 algorithm was applied to detect cervical cancer by classifying the patient's biopsy results based on attributes that affect the risk of cervical cancer from the dataset of a hospital in Caracas, Venezuela. The used of 858 data is imbalanced data conditions because there are dominating classes from other classes. Imbalanced data problems can cause the inaccurate classification. To overcome this problem, researchers applied the undersampling technique to the majority data so the data will be balanced and obtained accurate results.

The researcher conducted a test with 165 data. The testing with comparison of training data and descending data testing of 80:20 show the highest accuracy rate was 75.76%. Based on the results of the calculation it can be concluded that the C4.5 algorithm can be applied and find interesting classification rules from the formed decision tree.

Keyword: C4.5 algorithm, undersampling, cervical cancer.