ABSTRACT

APPLICATION OF K-NEAREST NEIGHBOR ALGORITHM TO CATEGORICAL DATA BY CALCULATING DISTANCE USING WEIGHTED SIMPLE MATCHING COEFFICIENT

By

CITRA PUSPA TRIA

K-Nearest Neighbor is one of classification algorithm that classifies objects based on the majority class of k-nearest objects. Commonly, the measure of proximity between objects is calculated using Euclidean distances. However, if the data being used is categorical, Euclidean distances are seen as improper to apply. The weighted simple matching coefficient (WSMC) method, which calculates how close two objects are to one another, is one solution to this issue. This research was conducted to build a classification model for categorizing laptop pricing ranges based on 4 features offered, namely laptop brand, processor, RAM capacity, and storage capacity. The classification model is built on two proportion of training data and testing data, which is 80:20 and 90:10. Based on the analysis's results, the model with a 90:10 data ratio was found to be the best classification model, with accuracy values of 86.96%, recall of 50%, precision of 66.67%, and f1-score of 57.14%.

Keywords: Classification, K-Nearest Neighbor (KNN), Weighted Simple Matching Coefficient (WSMC)

ABSTRAK

PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* PADA DATA KATEGORIK DENGAN PERHITUNGAN JARAK MENGGUNAKAN *WEIGHTED SIMPLE MATCHING COEFFICIENT*

Oleh

CITRA PUSPA TRIA

K-Nearest Neighbor merupakan salah satu algoritma klasifikasi yang mengklasifikasikan objek berdasarkan kelas mayoritas dari *k* objek terdekat di sekitarnya. Pada umumnya, ukuran kedekatan antar objek dihitung dengan menggunakan jarak Euclidean. Akan tetapi, jarak Euclidean dirasa kurang tepat digunakan apabila data yang digunakan adalah data kategorik. Salah satu alternatif dalam mengatasi permasalahan tersebut adalah dengan menggunakan metode *weighted simple matching coefficient* (WSMC) sebagai ukuran kedekatan antar objek. Penelitian ini dilakukan untuk membangun model klasifikasi untuk penentuan kelas rentang harga laptop berdasarkan 4 fitur yang ditawarkan, yaitu merek laptop, prosesor, kapasitas RAM, dan kapasitas penyimpanan. Model klasifikasi dibangun berdasarkan dua proporsi pembagian data latih dan data uji, yaitu 80:20 dan 90:10. Berdasarkan hasil analisis, diperoleh model klasifikasi terbaik, yaitu model dengan proporsi 90% data latih dan 10% data uji, dengan nilai akurasi sebesar 86.96%, *recall* sebesar 50%, presisi sebesar 66.67%, dan *f1-score* sebesar 57.14%.

Kata Kunci: Klasifikasi, K-Nearest Neighbor (KNN), Weighted Simple Matching Coefficient (WSMC)