Application of Recommendation Medical Specialty Doctors Based on User Symptoms Using the C4.5 Method and K-Nearest Neighbor

Hendri1*, Viny Christanti Mawardi2, Prof. Dali S. Naga3
1,2,3 Informatics Engineering Department, Faculty of Information Technology, Universitas Tarumanagara

*lasmana_hendri@yahoo.com, viny@untar.ac.id

Abstract. This paper is about web based application of recommendation medical specialty doctors based on user symptoms using the C4.5 method and K-Nearest Neighbor. Classification is the process of grouping things based on classes with the characteristics of similarities and differences. This objects or entities are labeled as user symptoms. User symptoms are classified and produce a category of doctor that matches the user's symptoms. This application will be using sample data from ascertain the opinions and experience of Doctor Yohannes Cahyadi who practices at the Griya Kasih Indah Clinic and from the website alodokter, hellosehat, docdoc.com, cicendoeeyehospital, and the eye clinic nusantara. The purpose of making this application is to choose the right specialist doctor based on human symptoms or medical need. The test results obtained said that the best classification accuracy is to use the K-Nearest Neighbor method with an accuracy value of 100%, this shows that the use of the C4.5 method and K-Nearest Neighbor can provide recommendations for the right doctor with user symptoms.

Keyword- Classification, Symptoms, C4.5, K-Nearest Neighbor

1. Introduction
Advances in information technology have increasingly developed rapidly in all fields of life, and one of them is the advancement of information technology in the health sector. Health is the most valuable treasure for every human being. For this reason, the need for access to health that is fast, practical and easily are essential by the people of Indonesia. Recently, applications in the health sector have emerged, including AloDokter application and HaloDoc application.

The HaloDoc and AloDokter applications do not yet have a feature to recommend users to specialist doctors based on user symptoms. Therefore making a computer program about the application of recommendation medical specialty doctors based on user symptoms is needed for the people of Indonesia. Because with this program, users don't need to consult a general practitioner to find out the right doctor category based on their symptoms.

This application is made using the C4.5 method and K-Nearest Neighbor. C4.5 algorithm is a group of Decision Tree algorithms. This algorithm has input in the form of training samples and samples. Training samples are sample data that is used to build a tree that has been tested for correctness. While samples are data fields which will be used as parameters in classifying data [1]. K-Nearest Neighbors Algorithm is a method for classification of patient symptoms to an object data by using the minimum distance approach of an object to another object. K-Nearest Neighbors is a supervised method which means that testing data must have an existing reference value to carry out the classification process [2]. By making this program it is expected to provide information to users to provide recommendations for medical specialty doctors based on user symptoms.

2. Methodology
The method used to build this application is the C4.5 method and K-Nearest Neighbor.

2.1. C4.5 Method
C4.5 algorithm is the development of the ID3 algorithm (Iterative Dichotomiser 3). A series of improvements made to the ID3 algorithm reach its peak by producing a practical and influential system for the formation of decision trees [3]. This fields are called as user symptom data. Following is an explanation of the work steps of the C4.5 algorithm [4]:

1. Calculate the entropy of each attribute
   Calculate the entropy of each attribute using the formula:
   \[\text{Entropy}(S) = - \sum_{i=1}^{n} p_i \log_2 p_i\] (1)

2. Calculate the information gain of each attribute.
   Calculate information gain using the formula:
   \[\text{Gain}(S,A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \text{Entropy}(S_i)\] (2)

3. Calculate the split information of each attribute
   Calculate the split information value of each attribute using the formula:
   \[\text{Split Info}(S,A) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times \log_2 \frac{|S_i|}{|S|}\] (3)

4. Calculate the gain ratio value of each attribute
   Calculate the gain ratio of each attribute using the formula:
   \[\text{Gain ratio}(S,A) = \frac{\text{Information Gain}(S,A)}{\text{Split Info}(S,A)}\] (4)

The purpose of calculating Information Gain and Split Information is to get the value of the gain ratio. Attributes in the Decision Tree can be determined based on the value of the largest to the smallest gain ratio.

![Flowchart Algorithm C4.5](image)

**Figure 1.** Flowchart Algorithm C4.5

### 2.2. K-Nearest Neighbor Method

Research on the K-Nearest Neighbor algorithm mostly discusses how to select and weight features, so that classification performance runs well [5]. This method are using user symptom data as data testing. Calculation of distances on K-Nearest Neighbors uses distance calculation with Euclidean distance with the formula:

\[d_{(x,y)} = \sqrt{(D_{Test_1} - D_{Training_1})^2 + \ldots + (D_{Test_n} - D_{Training_n})^2}\] (1)
3. Results

This program was created to choose the right specialist doctor based on human symptoms or medical need. This program can be used by users by entering 55 symptoms as shown in Figure 3. Some symptoms that can be chosen by the user are fever, toothache, disturbed olfaction, olfactory disrupted, ear buzzing, difficulty swallowing, dry cough, phlegm cough, coughing up blood, easy to feel tired, heart palpitations, breath sounds, shortness of breath, complaints in the eye, red eyes, decreased vision, eye pain, headache, foggy vision, watery eyes, rubbed eyes, swaying eyes, double vision, protruding eyes, uninterrupted eye nerves, tense eyes, frequent squinting, normal eye refraction, eyes swollen, itchy eyes, the eyes feel hot, there is a lump in the eye, eyelids fall, eyelids fold, eyelids flipped, eyes losing distance, eyes losing depth etc.

Data for this application were obtained from interviews with doctor Yohannes Cahyadi and data from the internet namely the website of alodokter, hellosehat, docdoc.com, cicendoeyehospital, and the eye clinic nusantara. The data used to build this application are 184 data, and 16 specialty of doctors.

The test results from the Application of Recommendation Medical Specialty Doctors Based on User Symptoms using the C4.5 and K Nearest Neighbor methods consist of two stages of testing, can be stated as follows:

3.1. Testing of Program Performance

The results of the classification of the C4.5 method from the system calculation if the testing data are of adult age (19 - <= 65 years) AND fever = fever AND lung = shortness of breath can be seen in Figure 4.
Here are the results of the classification of the $C4.5$ method from the manual calculation if the testing data are of adult age ($19 \leq 65$ years) AND fever = fever AND lung = shortness of breath can be seen in Figure 5.

Following are the classification results of the $K$-Nearest Neighbor method from the system calculation if the testing data is of adult age ($19 \leq 65$ years) AND fever = fever AND lung = shortness of breath can be seen in Figure 6.

The following are the results of the classification of the $K$-Nearest Neighbor method from the manual calculation if the testing data is of adult age ($19 \leq 65$ years) AND fever = fever AND lung = shortness of breath can be seen in Table 1.

| Peringkat Euclidean Distance | Klasifikasi Rekomendasi Dokter |
|-----------------------------|--------------------------------|
| 1                           | Dokter Paru                    |
| 2                           | Dokter Umum                    |
Subsequent testing is carried out with an evaluation to find out which method has a high degree of accuracy. The evaluation conducted is using a confusion matrix.

3.2. Confusion Matrix

Tests carried out on 20 patient data. The data is used to evaluate whether the C4.5 method has produced predictions of doctor recommendations that are in accordance with actual doctor recommendations or not. The results of confusion matrix C4.5 against 20 patient data can be seen in Figure 7.

![Confusion Matrix C4.5](image7)

Figure 7. Confusion Matrix C4.5

Predictions study of doctor's recommendations are precision 100%, Recall 100%, and accuracy 80%.

Following is the test for the K-Nearest Neighbor method. Tests carried out on 20 patient data. The data is used to evaluate whether the K-Nearest Neighbor method has produced predictions of doctor recommendations that are in accordance with actual doctor recommendations or not. The results of K-Nearest Neighbor's confusion matrix of 20 patient data can be seen in Figure 8.

![Confusion Matrix K-Nearest Neighbor](image8)

Figure 8. Confusion Matrix K-Nearest Neighbor
Predictions study of doctor's recommendations are precision 100%, recall 100%, and accuracy 100%.

| No | Accuracy | Method                |
|----|----------|-----------------------|
| 1  | 100%     | K-Nearest Neighbor    |
| 2  | 80%      | C4.5                  |

### 4. Conclusion

The conclusions that can be obtained based on the results of tests that have been conducted on this application of recommendation medical specialty doctors based on user symptoms utilizing the C4.5 method and K-Nearest Neighbor are as follows:

1. All features contained in the existing modules can run and function properly and can be used properly in accordance with existing specifications.
2. Test results on the classification of the C4.5 method show the results of the correct doctor's recommendation, and have an accuracy of 80%.
3. Test results on the classification of the K-Nearest Neighbor method show the results of the correct doctor's recommendation, and have 100% accuracy.
4. The best method between C4.5 and K-Nearest neighbor, the K-Nearest Neighbor method has 100% accuracy.

### References

[1] Selvia Lorena Br Ginting, Wendi Zarman, Ida Hamidah. “Analisis dan Penerapan Algoritma C4.5 dalam Data Mining Untuk Memprediksi Masa Studi Mahasiswa Berdasarkan Data Nilai Akademik.” Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST), 2014.

[2] Translation from Nobertus Krisandi, Helmi, Bayu Prihandono. (2014). ALGORITMA K-NEAREST NEIGHBOR DALAM KLASIFIKASI DATA HASIL PRODUKSI KELAPA SAWIT PADA PT. MINAMAS KECAMATAN PARINDU, jurnal.untan.ac.id/index.php/jbmstr/article/.../1484, 28 Februari 2014, Accessed on July 2020, 19th.

[3] Raharja, Yosoa Putra. “Rancangan Bagun Sistem Rekomendasi Beasiswa Menggunakan Algoritma Klasifikasi C4.5 Pada Universitas Dian Nuswantoro.” Skripsi Fakultas Ilmu Komputer, Undinus. Semarang: Universitas Dian Nuswantoro, 2014.

[4] Ernest, Michael. “Prediksi Masa Studi Mahasiswa Menggunakan Metode Decision Tree Dengan Algoritma C4.5 Pada Aplikasi Konsultasi Akademik Online.” Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Tarumanagara (Skripsi tidak dipublikasikan), 2017.

[5] Translation from Alfian Sukma, Dian Ramadhan. (2017). Algoritma K-Nearest Neighbor, https://informatikalagi.com/algoritma-k-nn-k-nearest-neighbor/, Accessed on July 2020, 19th.