JIMP : Jurnal Informatika Merdeka Pasuruan
Vol.6 No.3 Desember 2021, P-ISSN : 2502-5716, E-ISSN : 2503-1945

Terakreditasi Peringkat Sinta 4 berdasarkan Petikan dari Keputusan Menteri Riset dan Teknologi/ Kepala Badan Riset dan Inovasi Nasional Nomor 200/M/KPT/2020, masa berlaku mulai Vol.3 No.1 tahun 2018 s.d Vol.7 No.2 tahun 2022

# Pengelompokkan Judul Buku dengan Menggunakan Algoritma K-Nearest Neighbor (K-NN) dan Term Frequency – Inverse Document Frequency (TF-IDF)

Fahrur Rozi [1], Farid Sukmana[2], Muhammad Nabil Adani [3]
[1,3] Pendidikan Teknologi Informasi, Fakultas Sains dan Teknologi, Universitas Bhinneka PGRI, Indonesia
[2] Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Gresik, Indonesia
[1]fahrur@ubhi.ac.id
[2]faridsukmana@outlook.com
[3]nblid48@gmail.com

Received : 21-11-2021; Accepted: 31-12-2021; Published: 31-12-2022

*Abstrak*—***Perpustakaan Universitas Bhinneka PGRI memiliki banyak koleksi baik dalam bentuk tercetak maupun digital, yang mana koleksi tersebut akan bertambah dengan seiring waktu berjalan. Dengan demikian jumlah koleksi buku yang ada di perpustakan akan semakin banyak dan beragam maka akan mengakibatkan sulitnya proses pengelompokan koleksi-koleksi yang ada. Metode yang digunakan dalam penelitian ini adalah data mining dengan pendekatan algoritma K-Nearest Neighbors (K-NN) dengan mengkombinasikan TF-IDF sebagai pembobotan frekuensi kata. Adapun tahap pengerjaan metode K-NN dalam penelitian ini melalui 4 tahap yaitu : (1) text preprocessing dengan menerapkan metode tokenisasi, case folding, stopword removal dan stemming, (2) Pembobotan kata dengan metode TF-IDF (3). Pemodelan nilai k dari batas minimal 1 sampai batas maksimal 30. (4) Klasifikasi data menggunakan nilai k paling optimal berdasarkan pemodelan nilai k. (5) pembahasan hasil klasifikasi. Teknik pengumpulan data menggunakan studi kepustakaan dan dataset. Dengan sistem klasifikasi ini diharapkan dapat memberikan informasi bermanfaat bagi pengguna.Selain itu penelitian ini juga bertujuan mengimplementasikan metode K-NN dengan mengkombinasikan dengan TF-IDF sekaligus mengetahui akurasi yang dihasilkan sistem prediksi penjualan tersebut. Hasil dari penelitian ini berdasarkan nilai akurasi tertinggi terhadap klasifikasi judul buku sebesar 66.67% dan nilai akurais terendah sebesar 60% dengan rata-rata nilai akurasi sebesar 63.33%.***

***Kata kunci*— Data Mining, K-Nearest Neighbor (K-NN), TF-IDF**

*Abstract*— ***Universitas Bhinneka PGRI Library has many collections in both printed and digital forms, which collections will increase over time. Thus the number of collections of books in the library will be more and more diverse, it will make the process of grouping existing collections difficult. The method used in this study is data mining with the K-Nearest Neighbors (K-NN) algorithm approach by combining TF-IDF as word frequency weighting. The stages of working on the K-NN method in this study went through 4 stages, namely: (1) text preprocessing by applying the tokenization method, case folding, stopword removal and stemming, (2) Word weighting using the TF-IDF method (3). Modeling the k value from a minimum limit of 1 to a maximum limit of 30. (4) Classification of data using the most optimal k value based on k value modeling. (5) discussion of classification results. Data collection techniques using literature studies and datasets. With this classification system, it is expected to provide useful information for users. In addition, this study also aims to implement the K-NN method by combining it with TF-IDF while at the same time knowing the accuracy of the sales prediction system. The results of this study are based on the highest accuracy value for the classification of book titles of 66.67% and the lowest accuracy value of 60% with an average accuracy value of 63.33%.***

***Kata kunci*— Data Mining, K-Nearest Neighbor (K-NN), TF-IDF**

## I. PENDAHULUAN

Bagian rangka mencerdaskan kehidupan bangsa Indonesia diperlukan suatu pendidikan yang sesuai dengan kebutuhan masyarakatnya. Salah satu upaya yang dilakukan untuk mencerdaskan kehidupan bangsa yakni mendirikan perpustakaan sebagai sarana dan prasarana untuk mendukung tujuan tersebut, oleh karena itu perpustakaan tidak dapat dilepaskan dari perkembangan masyarakat saat ini. Kondisi perpustakaan juga mencerminkan peradaban dan kebudayaan suatu bangsa.

Saat ini revolusi perpustakaan sudah memasuki wilayah tulungagung, revolusi perpustakaan ini ditandainya dengan banyak perpustakaan online daerah yang bermunculan salah satu contohnya perpustakaan online Universitas Bhinneka PGRI yang bisa diakses melalui simperpus.ubhi.ac.id. Berdasarkan informasi yang penulis dapatkan dari saat ini saat ini jumlah anggota Perpuda Tulungagung yang meminjam buku tahun 2019, sebanyak 27.334 orang. Sedangkan kunjungan setiap harinya mencapai 200-230 orang. Jika ada kunjungan gabungan, angka pengunjung bisa mencapai 300 orang per hari .

Perpustkaan Universitas Bhinneka PGRI atau Perpustakaan GRAHA PUSTAKA Universitas Bhinneka merupakan perpustakaan utama di lingkungan Universitas Bhinneka PGRI. Perpustakaan ini menjadi referensi dalam bidang pendidikan dengan menyediakan akses informasi dan pengetahuan yang lengkap, baik dalam bentuk koleksi



DOI: http://dx.doi.org/10.37438/jimp.v6i3.346

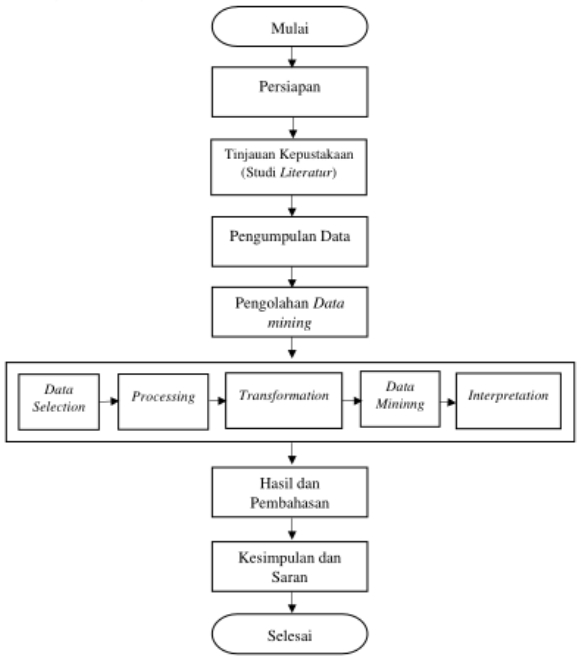
tercetak maupun digita. Perpustakaan ini dikelola dengan memanfaatkan teknologi informasi dan komunikasi dalam menunjang pelayanannya kepada pengguna yang bisa diakses secara online melalui simperpus.ubhi.ac.id dan offline oleh mahasiswa. Perpustakaan ini memiliki banyak koleksi baik dalam bentuk tercetak maupun digital, yang mana koleksi tersebut akan bertambah dengan seiring waktu berjalan. Dengan demikian jumlah koleksi buku yang ada di perpustakan akan semakin banyak dan beragam maka akan mengakibatkan sulitnya proses pengelompokan koleksi-koleksi yang ada, sehinggal akan lebih memudahkan pengguna apabila koleksi tersebut tersusun rapi dalam kelompoknya masing-masing agar memudahkan pembaca menemukan dan menggunakan koleksi tersebut yang secara otomatis telah disediakan oleh sistem.

Salah satu metode yang mampu untuk menangani dalam pegelompokan dokumen adalah K-NN. K-NN merupakan algoritma yang sering digunakan dalam proses klasifikasi suatu objek yang berdasarkan data training yang jaraknya paling dekat dengan objek tersebut. Nearest neighbor merupakan suatu pendekatan yang dilakukan untuk menghitung kedekatan antara kasus baru (data test) dan kasus lama (data training) yang berdasarkan pencocokan bobot dari sejumlah fitur yang ada. Beberapa penelitian yang menerapkan K-NN diantaranya [1], merupakan penelitian K-NN dalam penyakit diabetes. Penelitian lainnya yaitu [2] penggunaan K-NN dalam text classification. Selain itu untuk pembobotan serta pencarian token – token kata diperlukan sebuah metode yaitu Term Frequency – Inverse Document Frequency (TF-IDF). Salah satu penelitian mengenai metode Term Frequency – Inverse Document Frequency (TF-IDF) [3] yaitu mengenai pencarian relevansi dokumen berdasarkan query

Pada penelitian ini akan dilakukan pengelompokan dokumen dengan menerapkan algoritma K-NN serta menggabungkan dengan algoritma Term Frequency – Inverse Document Frequency (TF-IDF) untuk menentukan frekuensi relatif dari setiap kata-kata atau token-token yang mana setiap kata tersebut akan diberikan pembobotan. Penelitian ini diharapkan mampu memberikan solusi sebuah metode dalam pengelompokan dokumen atau judul buku diperpustakaan.

## II. METODOLOGI PENELITIAN

Penelitian ini memiliki alur penelitian dengan tahapan didalam proses Prepocessing data yatu : tokenisasi, case folding, stopword removal, dan stemming. Sementara tahapan berikutnya adalah algoritma Term Frequency – Inverse Document Frequency (TF-IDF) untuk menentukan frekuensi relatif dari setiap kata-kata atau token-token yang mana setiap kata tersebut akan diberikan pembobotan. Alur penelitian secara keseluruhan digambarkan pada gambar 1.




Gambar 1. Alur Penelitian

### A. Data Mining

Data Mining adalah serangkaian proses pengumpulan data melalui penemuan pola yang dilakukan secara sistematis terhadap sekumpulan data yang berdimensi besar, kemudian data yang telah dikumpulkan dapat digunakan untuk diolah lebih lanjut menjadi informasi yang bermanfaat. Data mining yang merupakan subbidang dari computer science juga dikenal sebagai Knowledge Discovery in Databases (KDD) [4]. KDD merupakan proses pencarian informasi data yang bermanfaat dari kumpulan data atau informasi yang ada. menyatakan "data mining merupakan serangkaian proses untuk menggali nilai tambah berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data." Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database besar . Tugas dari data mining sendiri melakukan analisis secara otomatis atau semi otomatis terhadap kumpulan data atau informasi yang berdimensi besar bertujuan untuk mencari pola-pola tertentu yang sebelumnya tidak diketahui oleh sistem.

### B. Text Mining

Text mining atau dikenal juga sebagai text data mining merupakan suatu varian dari data mining. Text mining bisa diartikan suatu proses pencarian informasi yang berkualitas tinggi yang diperoleh dari suatu kumpulan dokumen atau informasi berupa teks yang diperoleh dari prediksi pola-pola tertentu yang memiliki ciri yang khas melalui sarana seperti pembelajaran pola statistik. Text mining juga dapat diartikan sebagai suatu metode yang digunkana untuk mencari informasi baru yang tidak diketahui sebelumnya oleh komputer yang berjalan secara otomatis atau



DOI: http://dx.doi.org/10.37438/jimp.v6i3.346

semiotomatis untuk mengekstrak informasi dari berbagai sumber yang berbeda [5].

### C. Preprocessing Data

Preprocessing data merupakan suatu tahapan yang dilakukan sebelum pembelajaran pada data dimulai. Tujuan dari preprocessing ini yakni menghasilkan sebuah set term index yang dapat mewakili dari dokumen atau informasi yang ada. Preprocessing data merupakan suatu tehnik pembersihan data agar data dapat diolah yang menggunakan metode-metode yang ada pada machine learning. Metode-metode yang dapat digunaan pada tahap preprocessing data ini meliputi:

### D. Tokenisasi

Secara garis besar tokenisasi merupakan suatu metode untuk memecahkan atau memenggal sekumpulan karakter utuh yang berup kalimat, paragraph, atau dokumen menjadi satuan kata atau token-token. Proses ini akan memudahkan program untuk melakukan proses klasifikasi . Contoh dari proses ini yaitu "Perpustakaan Universitas Bhinneka PGRI berada di kabupaten Tulungagung" dipecah menjdi satuan kata atau token-menjadi "Perpustakaan", "Universitas", "Bhinneka", "PGRI", "berada", "di", "kabupaten", "Tulungagung". Untuk melakukan tokenisasi menggunakan bahasa pemrograman python bisa menggunakan fungsi yang telah tersedia pada library string. Contohnya str.split("perpustakaan universitas", " ") [6].

### E. Case Folding

Suatu proses pengubahan kata yang menggunakan huruf kapital menjadi huruf kecil serta menghilangkan semua tanda baca sehingga menyisakan kalimat yang berupa a sampai z saja. Untuk melakukan case folding menggunakan bahasa pemrograman python bisa menggunakan fungsi yang telah tersedia pada library string. Contohnya str.lower("Perpustakaan Universitas") maka menghasilkan "perpustakaan universitas"

### F. Stopword Removal

Suatu proses untuk menghilangkan kata-kata imbuhan pada kalimat serta kata-kata yang sering muncul namun tidak memiliki pengaruh apapu dalam ekstraksi data. Kata yang termasuk dalam list yang akan dihilangkan dalam proses stopword removal ini biasanya berupa kata penunjuk waktu dan kata tanya. Pada proses stopword removal ini penulis menggunakan external library dari bahasa pemrograman python yakni menggunakan library PySastrawi yang ditulis oleh Hanif Amal Robbani yang bisa ditemukan di https://github.com/har07/PySastrawi.[7]

### G. Stemming

Proses merubah kata yang memiliki imbuhan menjadi kata dasar. Metode pengubahan bentuk kata menjadi kata dasar menyesuaikan dengan struktur bahasa yang digunakan. Proses ini melibatkan beberapa algoritma seperti algorima porter, algoritma nazief-andriani yang umum digunakan untuk mengolah teks berbahasa Indonesia. Proses stemming pada teks berbahasa Indonesia cukup berbeda dengan stemming pada teks berbahasa inggris, pada teks berbahasa inggris proses yang diperlukan hanya menghilangkan sufiks. Sedangkan pada teks berbahasa indonesia perlu menghilangkan sufiks, prefix dan konfiks. Sebagai contoh stemming pada teks berbahasa indonesia yakni kata "membaca" menjadi "baca", "menulis" menjadi "tulis", dan "mencatat" menjadi "catat" .

### H. Term Frequency – Inverse Document Frequency (TF-IDF)

Algoritma Term Frequency – Inverse Document Frequency (TF-IDF) merupakan salah satu algoritma yang sangat populer untuk melakukan klasifikasi teks yang dapat digunakan untuk menganalisis hubungan antara sebuah kalimat dengan sekumpulan dokumen. Metode TF-IDF merupakan metode untuk menentukan frekuensi relatif dari setiap kata-kata atau token-token yang mana setiap kata tersebut akan diberikan pembobotan berupa nilai berdasarkan penting atau tidaknya suatu kata dalam dokumen berdasarkan jumlah kemunculan kata dalam suatu dokumen dan mengukur kata-kata tersebut terhadap keseluruhan dokumen yang ada. Metode TF-IDF ini merupakan gabungan dari dua dua konsep untuk penghitungan bobot yaitu Tf (term frequency) dan idf (inverse document frequency) [3][8].

Persamaan Tf (term frequency) dapat dilihat pada persamaan 1.

$$TF = \frac{t}{d} \quad (1)$$

Dimana t adalah jumlah kemunculan kata tertentu dalam dokumen d. Sementara d : total keseluruhan kata pada dokumen. Persamaan idf (inverse document frequency) dapat dilihat pada persamaan 2.

$$idf = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

Dimana N adalah total dokumen yang ada, serta df(t) : jumlah dokumen yang memiliki kata t. Persamaan TD-IDF dapat dilihat pada persamaan 3.

$$TFidf = TF.idf \quad (3)$$

### I. K-Nearest Neighbor

K-Nearest Neighbor(K-NN) termasuk dalam kelompok instance-based learning. Algoritma ini merupakan salah satu tehnik lazy-learning. K-NN merupakan algoritma yang sering digunakan dalam proses klasifikasi suatu objek yang berdasarkan data training yang jaraknya paling dekat dengan objek tersebut. Nearest neighbor merupakan suatu pendekatan yang dilakukan untuk menghitung kedekatan antara kasus baru (data test) dan kasus lama (data training) yang berdasarkan pencocokan bobot dari sejumlah fitur yang ada. Jika mengacu pada penelitian ini objek yang digunakan berupa teks. Algoritma ini nantinya akan menglasifikasikan suatu dokumen yang belum diketahui kategorinya terhadap dokumen latih yang telah diketahui



DOI: http://dx.doi.org/10.37438/jimp.v6i3.346

kategorinya dengan menggunakan rumus euclidean dalam dimensi sebesar k yang mengelilingi dokumen tes yang belum diketahui kategorinya, sehingga setelah jarak terhitung maka jarak yang paling dekat dengan dokumen latih dianggap memiliki kesamaan [9]. Dengan persamaan 3 euclidean distance sebagai berikut:

$$d = \sqrt{(q_1 - p_1)^2 + (q_1 - p_2)^2 + ...(q_n - p_n)^2} \quad (4)$$

Dimana q adalah kelas pertama, i adalah variable data, n dimensi data, p jarak kedua, serta d jarak[10].

## III. HASIL DAN PEMBAHASAN

### A. Data

Dalam penulisan ini, penulis menggunakan data teks berupa judul buku yang ada di perpustakaan Universitas Bhinneka PGRI. Data ini akan dibagi menjadi dua bagian untuk memodelkan dan menguji hasil pembelajaran algoritme klasifikasi yang penulis gunakan yaitu K-Nearest Neighbor (K-NN). Dengan data awal sebanyak 299 buah data yang dibagi menjadi data latih dan data uji dengan masing-masing pembagian yaitu 60% data latih dan 40% data uji untuk menentukan nilai k-nya nanti.

Data yang digunakan dalam penelitian ini memiliki 4 jenis kategori. Kategori ini akan digunakan sebagai label dalam pengelompokan jenis buku. Kategori buku dapat dilihat dalam tabel I.

TABEL I
ANALISIS KATEGORI

| No | Kategori | Jumlah |
|---|---|---|
| 1 | Hukum | 57 |
| 2 | Teknologi | 68 |
| 3 | Kesustraan | 80 |
| 4 | Ilmu Murni | 94 |

### B. Pembersihan Judul Teks

Sebelum mengklasifikasikan kategori pada teks yang berupa judul buku, penulis melakukan preprocessing pada teks tersebut dengan beberapa metode. Tujuannya adalah untuk membersihkan kata-kata yang tidak diperlukan serta merubah kata-kata tersebut menjadi vektor larik sehingga dapat diproses oleh numerikal komputasi

#### 1) Case Folding

Proses ini untuk mengubah kata yang menggunakan huruf kapital menjadi huruf kecil serta menghilangkan semua tanda baca. Perbandingan antara judul buku sebelum dan sesudah case folding dapat dilihat pada tabel II.

TABEL II
PERBANDINGAN SEBELUM DAN SESUDAH CASE FOLDING

| No | Judul Buku | Keterangan |
|---|---|---|
| 1 | Berpikir Matematis : Matematika Untuk Semua | Sebelum |
| | berpikir matematis matematika untuk semua | Sesudah |
| 2 | Kancil Mencuri Ketimun | Sebelum |
| | kancil mencuri ketimun | Sesudah |
| 3 | Dasar algoritma dan struktur data dengan bahasa java + CD | Sebelum |
| | dasar algoritma struktur data bahasa java cd | Sesudah |

#### 2) Stopword Removal

Proses ini untuk menghilangkan kata-kata penghubung, beberapa contoh kata penghubung seperti ada, pada, dan, seseorang, sesuatu, tempat, kata dan sebagainya. Perbandingan antara sebelum dan sesudah case folding dapat dilihat apada tabel II.

TABEL II
PERBANDINGAN SEBELUM DAN SESUDAH CASE FOLDING

| No | Judul Buku | Keterangan |
|---|---|---|
| 1 | berpikir matematis matematika untuk semua | Sebelum |
| | berpikir matematis matematika untuk semua | Sesudah |
| 2 | kancil mencuri ketimun | Sebelum |
| | kancil mencuri ketimun | Sesudah |
| 3 | dasar algoritma dan struktur data dengan bahasa java cd | Sebelum |
| | dasar algoritma struktur data bahasa java cd | Sesudah |

#### 3) Stemming

Stemming adalah proses untuk mengubah kata berimbuhan menjadi kata dasarnya. Dari tabel II kalimat yang telah mengalami stopword removal dilanjutkan dengan steeming sehingga menjadi seperti tabel III.

TABEL III
JUDUL BUKU SETELAH DILAKUKAN STEMMING

| No | Judul Buku | Keterangan |
|---|---|---|
| 1 | berpikir matematis matematika untuk semua | Sebelum |
| | pikir matematis matematika untuk semua | Sesudah |
| 2 | kancil mencuri ketimun | Sebelum |
| | kancil curi timun | Sesudah |
| 3 | dasar algoritma dan struktur data dengan bahasa java cd | Sebelum |
| | dasar algoritma struktur data bahasa java cd | Sesudah |

### C. TFIDF

Dalam proses ini penulis akan menilai nilai kemunculan suatu kata dan menghitung bobotnya, tujuannya untuk melihat sesering apa kata tersebut muncul pada tiap-tiap dokumen. tabel IV hasil dari perhitungan TFIDF menggunakan persamaan 4.

TABEL IV
HASIL PERHITUNGAN TFIDF

| No | Term/Kata | Dokumen 1 | Dokumen 2 | Dokumen 3 |
|---|---|---|---|---|
| 1 | pikir | 0.4771 | 0 | 0 |
| 2 | matematis | 0.4771 | 0 | 0 |
| 3 | matematika | 0.4771 | 0 | 0 |
| 4 | untuk | 0.4771 | 0 | 0 |
| 5 | semua | 0.4771 | 0 | 0 |
| 6 | kancil | 0 | 0.4771 | 0 |
| 7 | curi | 0 | 0.4771 | 0 |
| 8 | timun | 0 | 0.4771 | 0 |

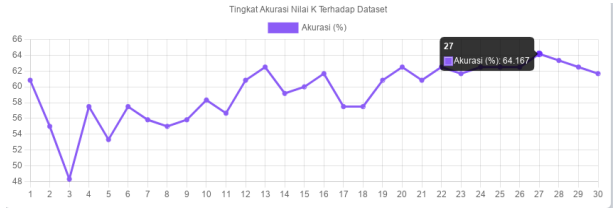


DOI: http://dx.doi.org/10.37438/jimp.v6i3.346

| | | | | |
|---|---|---|---|---|
| 9 | dasar | 0 | 0 | 0.4771 |
| 10 | algoritma | 0 | 0 | 0.4771 |
| 11 | struktur | 0 | 0 | 0.4771 |
| 12 | data | 0 | 0 | 0.4771 |
| 13 | bahasa | 0 | 0 | 0.4771 |
| 14 | java | 0 | 0 | 0.4771 |
| 15 | cd | 0 | 0 | 0.4771 |

### *D. Pemodelan Nilai K*

Pada algoritma K-NN terdapat nilai k yaitu nilai atau banyaknya jumlah tetangga terdekat yang sangat bergantung pada bagaimana bentuk data, seringkali dijumpai bahwa nilai k yang berbeda akan memberikan hasil akurasi yang berbeda pula, sehingga penulis membuat suatu model dengan mensimulasikan pengklasifikasian menggunakan nilai k yang selalu bertambah dengan limit bawah yaitu 1 hingga batas atas bernilai 30. Untuk mensimulasikan penyetelan nilai k terbaik, penulis membagi data menjadi dua bagian, sebanyak 299 data akan dibagi menjadi 60% data latih dan 40% data uji. Data ini nantinya akan dimasukkan kedalam simulasi penentuan nilai k yang akan menghasilkan grafik akurasi seperti pada gambar 2.




Gambar 2. Grafik tingkat akuras nilai k

Dari hasil simulasi gambar 2, dapat disimpulkan model dan akurasi terbaik berada pada saat iterasi ke 27 dengan nilai akurasi mencapai 0.64 atau 64%, dengan kata lain nilai k terbaik adalah 27.

### *E. Klasifikasi*

Penelitian ini bertujuan untuk mengklasifikan data judul buku menggunakan algoritme K-NN ke dalam 4 kategori yaitu hukum, ilmu murni, teknologi, kesustraan. Klasifikasi ini menggunakan perhitungan jarak Euclidean dengan hasil pengujian nilai k pada tabel V.

TABEL V
HASIL PENGUJIAN NILAI K TERHADAP DATA UJI

| **Data Uji (%)** | **Data Latih (%)** | | | **Keterangan** |
|---|---|---|---|---|
| | **60** | **70** | **80** | |
| 40 | 63.33 | 0 | 0 | cukup baik |
| 30 | 0 | 66.67 | 0 | Cukup baik |
| 20 | 0 | 0 | 60 | Cukup baik |

Simulasi pengujian nilai k yang dimulai dari 1 sampai 30 terhadap dataset menghasilkan k dengan nilai tertinggi sebesar 27 dengan akurasi 64.16% , nilai k yang dihasilkan dari proses simulasi digunakan untuk pengujian terhadap data uji.

## IV. KESIMPULAN

Judul buku yang berupa teks terbatas pada jumlah karakter, ataupun kata atau dengan kata lain pesan pendek, dapat diklasifikasi dengan baik akan tetapi dikarenakan jumlah data yang tidak seimbang mempengaruhi hasil dari klasifikasi yang telah di uji coba oleh penulis. Percobaan dengan mensimulasikan nilai k terlebih dahulu, pemodelan yang dibuat dapat menghasilkan nilai akurasi yang baik saat k bernilai 27 yaitu mencapai akurasi 64.16% terhadap dataset, setelah diuji sebanyak 3 kali dengan data uji yang berbeda menghasilkan tingkat akurasi rata rata sebesar 63.33%: Rendahnya tingkat akurasi pada penelitian ini disebabkan oleh ketidak seimbangan dataset yang dimiliki penulis yakni untuk topik hukum sebanyak 57, topik ilmu murni sebanyak 94, kesusastraan sebanyak 80, dan teknologi sebanyak 68. dari keempat topik tersebut topik ilmu murni dan kesustraan terlalu mendominasi dataset.

## REFERENSI


[1] A. Ali, M. Alrubei, L. F. M. Hassan, M. Al-Ja'afari, and S. Abdulwahed, "Diabetes classification based on KNN," *IIUM Eng. J.*, 2020, doi: 10.31436/iiumej.v21i1.1206.

[2] Z. Chen, L. J. Zhou, X. Da Li, J. N. Zhang, and W. J. Huo, "The Lao text classification method based on KNN," in *Procedia Computer Science*, 2020, doi: 10.1016/j.procs.2020.02.053.

[3] M. A. Rofiqi, A. C. Fauzan, A. P. Agustin, and A. A. Saputra, "Implementasi Term-Frequency Inverse Document Frequency (TF-IDF) Untuk Mencari Relevansi Dokumen Berdasarkan Query," *Ilk. J. Comput. Sci. Appl. Informatics*, 2019, doi: 10.28926/ilkomnika.v1i2.18.

[4] J. Hu, H. Peng, J. Wang, and W. Yu, "kNN-P: A kNN classifier optimized by P systems," *Theor. Comput. Sci.*, 2020, doi: 10.1016/j.tcs.2020.01.001.

[5] D. Ö. Şahin and E. Kılıç, "Two new feature selection metrics for text classification," *Automatika*, 2019, doi: 10.1080/00051144.2019.1602293.

[6] F. Rozi and F. Sukmana, "Document grouping by using meronyms and type-2 fuzzy association rule mining," *J. ICT Res. Appl.*, vol. 11, no. 3, 2017, doi: 10.5614/itbj.ict.res.appl.2017.11.3.4.

[7] S. Xu, "Bayesian Naïve Bayes classifiers to text classification," *J. Inf. Sci.*, 2018, doi: 10.1177/0165551516677946.

[8] F. Sukmana and F. Rozi, "Extraction keyterm in work order for decision support," *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 22, pp. 3262–3272, 2019.

[9] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *Int. J. Res. Mark.*, 2019, doi: 10.1016/j.ijresmar.2018.09.009.

[10] P. J. S. Ferreira, J. M. P. Cardoso, and J. Mendes-Moreira, "KNN prototyping schemes for embedded human activity recognition with online learning," *Computers*, 2020, doi: 10.3390/computers9040096.







DOI: http://dx.doi.org/10.37438/jimp.v6i3.346