SYSTEMIC: Information System and Informatics Journal

ISSN: 2460-8092, 2548-6551 (e)

Vol 6 No 2 – Desember 2020

Klasifikasi Text Judul Buku Perpustakaan Untuk Menentukan Kategori Buku Menggunakan K-Nearest Neighbor

Muhamad Kadafi

Universitas Islam Negeri (UIN) Raden Fatah Palembang

kadafi_uin@radenfatah.ac.id

Kata Kunci

Nearest Neighbor Classifier, Data mining, Perpustakaan.

Abstrak

Kebutuhan terhadap informasi dalam bentuk buku ataupun artikel ilmiah pada Perpustakaan UIN Raden Fatah Palembang semakin terus meningkat, untuk mempermudah dalam pencarian informasi buku salah satunya adalah dengan mengelompokan buku berdasarkan jenis kategorinya. Dalam mengelompokan data buku perpustakaan Metode Nearest Neighbor Classifier pada data mining dapat di kombinasikan dengan teknik ekstraksi data text untuk melakukan klasifikasi data text judul buku perpustakaan. Tujuan dari penelitian ini adalah untuk mengklasifikasikan text judul buku perpustakaan dengan menggunakan Nearest Neighbor Classifier untuk menentukan jenis kategori buku. Metode penelitian ini menggunakan teknik klasifikasi data mining Nearest Neighbor Classifier. Hasil dari penelitian ini adalah nilai akurasi tertinggi terdapat pada K=12 yaitu sebesar 72.50%, dan model yang terbentuk dapat digunakan untuk mengklasifikasikan buku dengan label 2x0, 150, 2x2, 400, 020, 2x1, 657, 500, 375, 302.2, 800 dan tidak dapat di gunakan untuk klasifikasi buku dengan label kelas 070, 370, 330, 300, 600, 340, 700.

Keywords

Nearest Neighbor Classifier, Data mining, Library

Abstract

The need for information in the form of books or scientific articles at the Library of UIN Raden Fatah Palembang continues to increase. To make it easier to find book information, one of which is by classifying books based on the type of category. In classifying library book data, the Nearest Neighbor Classifier method in data mining can be combined with text data extraction techniques to classify library book title text data. The purpose of this study was to classify the text title of library books using the Nearest Neighbor Classifier to determine the type of book category. This research method uses the Nearest Neighbor Classifier data mining classification technique. The results of this study are that the highest accuracy value is found at K = 12, which is 72.50%, and the model formed can be used to classify books with labels 2x0, 150, 2x2, 400, 020, 2x1, 657, 500, 375, 302.2, 800. and cannot be used for classifying books with class labels 070, 370, 330, 300, 600, 340, 700.

1. Pendahuluan

Universitas Islam Negeri (UIN) Raden Fatah Pelembang adalah salah satu perguruan tinggi negeri di kota Palembang Sumatera Selatan. UIN Raden Fatah Pelembang dilengkapi dengan berbagai fasilitas untuk menunjang kegiatan akademik, dimana salah satunya adalah fasilitas berupa perpustakaan. Perpustakaan adalah fasilitas atau tempat menyediakan sarana bahan bacaan. Tujuan dari perpustakaan khususnya perguruan tinggi adalah memberikan layanan informasi untuk kegiatan belajar, penelitian, dan pengabdian masyarakat dalam rangka melaksanakan Tri Dharma Perguruan Tinggi [1].

Perpustakaan UIN Raden Fatah Palembang memiliki koleksi buku yang cukup banyak dari berbagai cabang ilmu pengetahuan. Koleksi buku yang tersedia dapat digunakan oleh dosen, mahasiswa dan staf untuk mencari referensi buku atau sebagai bahan bacaan. Pada Perpustakaan UIN Raden Fatah Palembang kebutuhan terhadap informasi dalam bentuk buku ataupun artikel ilmiah semakin meningkat, hal tersebut dapat dilihat dari banyaknya jumlah buku yang terdapat pada perpustakaan. Maka untuk membantu mempermudah dalam pencarian informasi buku

DOI : 10.29080/systemic.v6i2.1056

salah satunya adalah dengan mengelompokan buku berdasarkan jenisnya. Pengelompokan buku ini dibutuhkan untuk mempermudah pencarian informasi yaitu berupa topik yang menggambarkan pokok pembahasan secara umum [2].

Dalam mengelompokan data buku perpustakaan para pengambil keputusan bisa memanfaatkan gudang data yang sudah dimiliki untuk melakukan analisa dalam mengambil keputusan, salah satunya yaitu dengan cara penggalian informasi atau pola yang penting dan menarik dari data jumlah besar, yang disebut dengan data mining. Data mining diartikan sebagai menambang data atau upaya untuk menggali informasi yang berharga dan berguna pada database yang sangat besar.

Salah satu teknik yang digunakan untuk menentukan pola tersebut yaitu dengan menggunakan teknik klasifikasi pada data mining dengan metode Nearest Neighbor Classifier. Nearest Neighbor Classifier adalah suatu metode yang menggunakan algoritma supervised [3], [4] , [5]. Supervised learning bertujuan untuk menemukan pola baru dalam data dengan menghubungan pola data yang sudah ada dengan data baru. Tujuan dari Nearest Neighbor Classifier adalah untuk mengklasifikasi objek baru berdasarkan atribut dan traning samples[4],[5]. Dimana hasil dari uji sampel yang baru di klasifikasikan berdasarkan mayoritas dari kategori pada ketetanggaan terdekat. Nearest Neighbor Classifier adalah metode yang menentukan nilai jarak pada pengujian data testing dengan data training berdasarkan nilai terkecil dari ketetanggan terdekat [6].

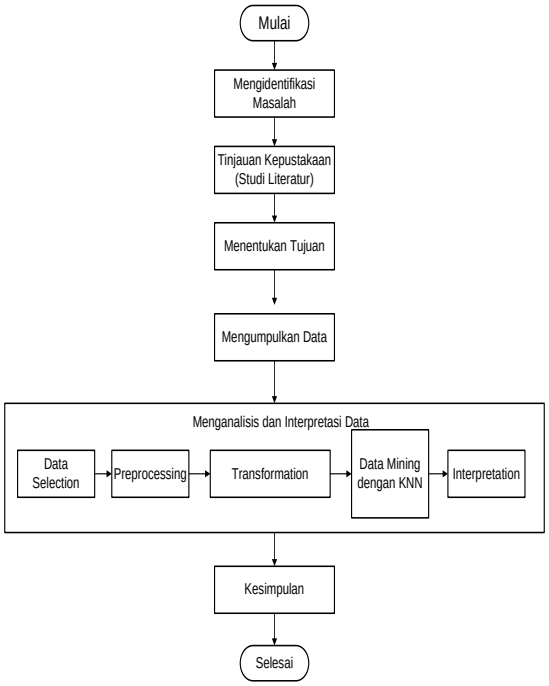
Untuk mengklasifikasikan buku perpustakaan, data buku yang telah tersedia sebelumnya dapat digunakan sebagai data training yaitu berupa data teks judul buku perpustakaan beserta label klasifikasinya. Nearest Neighbor Classifier pada data mining dapat di kombinasikan dengan teknik ekstraksi data text untuk melakukan klasifikasi dokumen, dalam hal ini adalah text judul buku perpustakaan yang disebut dengan text mining. Text mining merupakan proses menambang data yang berupa text dimana sumber data biasanya di dapat dari dokumen dan tujuannya adalah untuk mencari kata – kata yang dapat mewakili isi dari dokumen sehingga dapat dilakukan analisa keterhubungan antar dokumen [7]. Tujuan dari text mining adalah mengekstrak informasi yang berguna dari sumber data, sumber data yang digunakan pada text mining adalah sekumpulan dokumen yang memiliki format yang tidak terstruktur melalui indentifikasi dan eksplorasi pola yang menarik. Adapun tugas khusus dari text mining antara lain, pengkategorian text (text categorization) dan Pengelompokan Text (text clustering) [3].

2. Metode Penelitian

Metode penelitian yang digunakan adalah pendekatan kuantitatif karena proses pengolahan data yang akan digunakan pada penelitian ini bersifat kuatitatif, digunakan untuk menghitung dan mengukur pada saat analisis dan interpretasi data. Dan hasilnya berupa data prediksi klasifikasi kategori buku jenis penelitian yang digunakan adalah penelitian eksperimen.

2.1 Tahapan Penelitian

Adapun Tahapan Penelitian adalah seperti pada gambar 1:



Gambar 1. Tahapan Penelitian

Tahapan yang akan ditempuh yaitu:

1. Mengidentifikasi Masalah

Masalah yang diidentifikasi dalam penelitian ini adalah bagaimana mengkategorikan buku berdasarkan pada text judul buku perpustakaan dengan menggunakan Nearest Neighbor Classifier?

2. Tinjauan Kepustakaan (Studi Literatur)

Literatur – literatur yang di pakai sebagai bahan refensi dalam penelitian ini adalah dari jurnal – jurnal ilmiah, modul pembelajaran, dan buku tentang data mining. literatur – literatur ini akan menjadi pedoman untuk melakukan penelitian agar memudahkan proses penelitian.

3. Menentukan Tujuan

Tujuan pada penelitian ini adalah mengklasifikasikan text judul buku perpustakaan dengan menggunakan Nearest Neighbor Classifier untuk menentukan kategori buku.

4. Mengumpulkan Data

Metode pengumpulan data dilakukan dengan cara melakukan pengamatan langsung di

Perpustakaan UIN Raden Fatah Palembang. Teknik yang digunakan adalah teknik analisis dokumen yaitu data transaksi peminjaman buku perpustkaan tahun 2015, 2016, 2017.

5. Menganalisis dan Interprestasi Data

Tahapan ini merupakan pengolahan data mining dengan mengikuti tahapan KDD (Knowledge Discovery In Database), dan metode yang digunakan untuk proses pengolahan data mining menggunakan klasifikasi Nearest Neighbor.

6. Kesimpulan

Membuat kesimpulan dari hasil penelitian dan memberikan saran untuk pihak perpustakaan agar dapat menjadi lebih baik lagi.

2.2 Nearest Neighbor Classifier

Metode Nearest Neighbor Clasiffier pertama kali di jelaskan pada awal tahun 1950-an. Nearest Neighbor classifier di dasarkan pada pembelajaran dengan analogi, yaitu dengan membandingkan tuple uji yang diberikan dengan tuple pelatihan yang mirip dengannya. Tuple pelatihan di deskripsikan dengan n atribut. Setiap tuple merepresentasikan suatu titik dalam ruang berdimensi n. Dengan cara ini, semua tuple pelatihan disimpan dalam ruang pola berdimensi n. Ketika diberi tuple yang tidak diketahui K-NN Classifier mencari ruang pola untuk K tuple pelatihan yang paling dekat dengan tuple yang tidak diketahui. Tuple pelatihan K ini adalah K "Ketetanggan Terdekat" dari tuple yang tidak diketahui.

"Kedekatan" di definisikan dalam metrik jarak, misalnya seperti jarak Euclidean. Jarak Euclidean antara dua titik atau tuple, misal $\boldsymbol{X}_1$ = ($\boldsymbol{x}_{11}$, $\boldsymbol{x}_{12}$, … , $\boldsymbol{x}_{1n}$) dan $\boldsymbol{X}_2$ = ($\boldsymbol{x}_{21}$, $\boldsymbol{x}_{22}$, … , $\boldsymbol{x}_{2n}$), adalah

$$dist(\boldsymbol{X}_1, \boldsymbol{X}_2) = \sqrt{\sum_{i=i}^{n}(\boldsymbol{x}_{1i}, x_{2i})^2} \quad \ldots\ldots (1)$$

Dengan kata lain untuk setiap atribut numerik, kita mengambil selisih antara nilai yang

y. nilai cosinus 0 berarti kedua vector berada pada 90 derajat satu sama lain (orthogonal) dan tidak memiliki kecocokan. Semakin dekat nilai cosinus ke 1, semakin kecil sudutnya dan semakin besar kecocokan antar vector. Ukuran kesamaan kosinus disebut sebagai ukuran non metrik.

2.4 TF - IDF

TF-IDF adalah singkatan dari istilah term frequency – inverse document frequency. Term frequency adalah berapa kali term tersebut muncul dalam dokumen. Document frequency adalah istilah untuk jumlah dokumen yang berisi term tertentu. Persamaan untuk nilai – nilai ini ditunjukan pada gambar persamaan berikut [8] :

$$tf.idf(t,d) = tf(t,d).idf(t)$$

sesuai dari atribut pada tuple $\boldsymbol{X}_1$ dan di tuple $\boldsymbol{X}_2$, mengkuadratkan dan mengakumulasinya. Akar kuadrat jumlah total jarak yang terakumulasi. Biasanya, kita menormalkan setiap atribut sebelum menggunakan persamaan (1). Ini membantu mencegah atribut dengan rentang awalnya besar melebihi atribut dengan rentang awalnya lebih kecil. Normalisasi Min – maks, misalnya, dapat digunakan untuk mengubah nilai v dari atribut numerik A menjadi v' dalam rentang [0,1] dengan menghitung.

$$v' = \frac{v - min_A}{max_A - min_A}, \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Dimana min_A dan max_A adalah nilai minimum dan maksimum dari atribut A.

Untuk klasifikasi Nearest Neighbor, tuple yang tidak diketahui diberi kelas yang paling umum di antara k ketetanggan terdekat nya. Ketika k = 1, tuple yang tidak diketahui diberikan kelas dari tuple pelatihan yang paling dekat dengannya dalam ruang pola [5].

2.3 Cossine Similarity

Cossine similarity adalah ukuran kemiripan yang dapat digunakan untuk membandingkan dokumen atau memberikan peringkat dokumen sehubungan dengan vector dari query kata yang diberikan. Misalkan x dan y adalah dua vector untuk perbandingan. Menggunakan ukuran kosinus sebagai fungsi kesamaan.

$$sim(x,y) = \frac{x\,.y}{||x||\,||y||}, \ldots\ldots\ldots (3)$$

Dimana ||x|| adalah norma Euclidean dari vector x = $(x_1, x_2, \ldots, x_p)$ di definisikan sebagai $\sqrt{X_1^2 + X_2^2 + \ldots . + X_p^2}$ secara konseptual, ini adalah Panjang vector. Demikian pula ||y|| adalah norma Euclidean dari vaktor y. ukuran tersebut menghitung cosinus dari sudut antara vector x dan

$$tf(t,d) = \sum_{i \in d}^{|d|} 1\{d_1 = t\}, \ldots\ldots\ldots (4)$$

$$idf(t) = \text{Log}\left(\frac{|D|}{\sum_{d \in D}^{|D|} | t \in d}\right)$$

2.5 Evaluasi Model Klasifikasi

Evaluasi terhadap suatu klasifier umunya di lakukan menggunakan sebuah himpunan data uji, yang tidak digunakan dalam pelatihan classifier tersebut, dengan suatu ukuran tertentu. Terdapat sejumlah ukuran yang dapat digunakan untuk menilai atau mengevaluasi model klasifikasi, di antaranya adalah : accuracy atau tingkat pengenalan, error rate atau tingkat kesalahan atau kekeliruan klasifikasi, recall atau sensitivity atau true positive rate, specificity atau true negative

rate, precision, F-measure atau F_1 atau F-Score atau rata –rata harmonic dari precision dan recall, dan F_β (J. Han et al.2012), yang secara ringkas ilustrasikan pada Tabel 1.

Tabel 1. Ukuran Evaluasi Klasifikasi

No	Ukuran	Rumus
1.	Accuracy atau tingkat pengenalan	$\frac{TP+TN}{P+N}$
2.	Error rate atau tingkat kesalahan atau kekeliruan klasifikasi	$\frac{FP+FN}{P+N}$
3.	Recall atau sensitivity atau true positive rate	$\frac{TP}{P}$
4.	Specificity atau true negative rate	$\frac{TN}{N}$
5.	Precision	$\frac{TP}{TP+FP}$
6.	F atau F_1 atau F-Score atau rata – rata harmonic dari precision dan recall	$\frac{2\ x\ precision\ x\ recall}{precision + recall}$
7.	F_β, dimana β adalah sebuah bilangan rill nonnegatif	$\frac{1+\beta^2\ \ x\ precision\ x\ recall}{\beta^2\ x\ precision + recall}$

empat istilah sangat penting untuk memahami semua ukuran evaluasi dalam tabel tersebut adalah sebagai berikut :

1. TP atau True Positives adalah jumlah tuple positif yang dilabeli benar oleh classifier. Yang dimaksud tuple positif adalah tuple actual yang berlabel positif, seperti tuple dengan label Bonus = 'Ya'.
2. TN atau True Negative jumlah tuple negative yang dilabeli dengan benar oleh classifier. Yang dimaksud tuple negative adalah tuple actual yang berlabel negative, seperti tuple dengan label Bonus = 'Tidak'.
3. FP atau False Positives adalah jumlah tuple negative yang salah dilabeli oleh classifier. Misalnya, sebuah tuple pelanggan yang berlabel Bonus = 'Tidak' tetapi dilabeli Bonus = 'Ya'
4. FN atau False Negatives adalah jumlah tuple positif yang salah dilabeli oleh classifier. Misalnya, sebuah tuple pelanggan yang berlabel Bonus = 'Ya' tetapi oleh classifier di labeli dengan Bonus = 'Tidak'.

3. Hasil dan Pembahasan

Adapun tahapan yang digunakan dalam pengolahan data mining pada penelitian ini, yaitu mengikuti tahapan *Knowledge Discovery in Database* (KDD). Adalah sebagai berikut :

3.1 Data Selection

Data yang digunakan adalah data transaksi peminjaman buku tahun 2015, 2016, 2017. Transaksi Data peminjaman buku pada tahun 2015,2016, 2017 ada sebanyak 3239 Buah transaksi. Dibawah ini merupakan contoh sebagian dari data transaksi peminjaman buku.

Pada tahapan ini penulis hanya menggunakan 2 atribut yang nanti akan digunakan untuk proses pengolahan data mining. Yaitu :

1. Judul Buku.
   Berisi informasi judul buku.
2. Klasifikasi
   Klasifikasi adalah attribut yang berisi informasi kode jenis buku, penggolongan buku berdasarkan jenis buku.

3.2 Preprocessing

Pada tahapan ini *preprocessing* ini akan dilakukan proses integrasi data untuk menghubungkan tabel data peminjaman, selanjutnya dilakukan data *cleaning* untuk menghasilkan *dataset* yang bersih sehingga dapat digunakan dalam tahap berikutnya yaitu mining. Berikut merupakan penjelasan dari kedua proses tersebut yaitu:

1. Integrasi Data
   Tahap ini adalah proses penggabungan data dari berbagai *database* yang berbeda, sehingga data tersebut saling berintegrasi. Data integrasi dilakukan pada atribut - atribut yang mengidentifikasiskan entitas-entitas yang unik. Pada tahapan ini tidak ada penggabungan data dikarenakan data yang diambil berasal dari satu *database*.
2. Data *Cleaning*
   Pada tahapan ini data yang tidak *relevan, missing value,* dan *redudant* harus dibersihkan. Hal ini dikarenakan data yang *ralevan,* tidak *missing value,* dan tidak *redudant* merupakan syarat awal dalam

melakukan data mining. Suatu data dikatakan *missing value* jika terdapat atribut dalam *dataset* yang tidak berisi nilai atau kosong, sedangkan data dikatakan *redudant* jika dalam satu *dataset* lebih dari satu *record* yang berisi nilai yang sama.

3.3 Transformation

Tahap Transformation merupakan tahapan merubah data kedalam bentuk yang sesuai untuk selanjutnya di proses dalam pengolahan data mining. Berikut pada gambar 2 adalah sebagian transformasi data transaksi buku 2015, 2016, 2017.

	A	B
1	text	label
2	fiqh manajemen zakat di indonesia	2x0
3	jurnalistik hukum komunikasi massa jangkau era cyber communication milenium tiga	070
4	toward easy learning kiat sukses ajar di guru tinggi	370
5	10 jurus larang kok masih mau bisnis cara biasa	330
6	110 soal iman yang sehat akal	2x0
7	15 masalah fikih yang hangat kontroversial	2x0
8	20 salah dalam didik anak	370
9	22 nasihat abadi halus budi	300
10	25 pengaruh jiwa dan akal anak	150
11	254 hadits qudsi	2x2
12	40 hadis nabi saw telaah imam khomeini atas hadis mistis dan akhlak	2x2
13	5 jam ajar olah data dengan spss17	600
14	530 hadits sahih bukhari - muslim	2x2
15	65 manusia langit jalan hidup sahabat rasulullah saw	2x0
16	7 langkah awal tuju karier idam	300
17	81 putus hukum rasulullah saw	2x0
18	9 presentasi kreatif dengan powerpoint 2007	600
19	95 strategi ajar multiple intelligences	600
20	a concise introduction to linguistics	400
21	a practical english grammar	400
22	a successful foreign language guide	400

Gambar 2. Transformasi Data Peminjaman Buku

Berdasarkan gambar 2 di atas hasil dari transformasi data setelah melewati tahapan *data selection*, dan *preprocessing* menghasilkan Data Buku sebanyak 2.742 Judul buku serta klasifikasinya, yang terdiri 2 atribut yaitu:

1. Text : Atribut yang berisi informasi judul buku.
2. Label : Atribut yang berisi informasi klasifikasi buku.

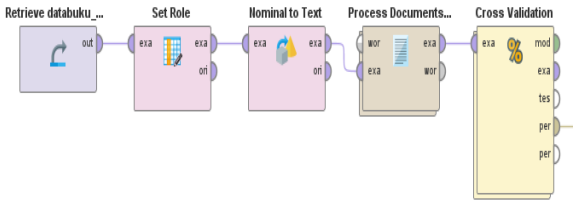
Jumlah buku yang terklasifikasi sebanyak 18 kelas. Berikut daftar buku yang klasifikasi. Pada Tabel 1 adalah daftar distribusi kelas buku.

Tabel 2. Daftar Distribusi Buku

No.	kelas	label kelas	jumlah
1	2x0	Agama Islam	755
2	070	Pers, Jurnalisme, Penerbitan, Persuratkabaran	18
3	370	Pendidikan	271
4	330	Ilmu Ekonomi	139
5	300	Ilmu Sosial	559
6	150	Psikologi	113
7	2x2	Hadist dan yang berkaitan	71
8	600	Teknologi (Ilmu Terapan)	198
9	400	Bahasa	86
10	020	Ilmu Perpustakaan dan Informasi	37
11	2x1	Alquran dan Ilmu yang berkaitan	143
12	657	Akuntansi	23
13	500	Sains dan Matematika	47
14	340	Ilmu Hukum	183
15	375	Kurikulum	15
16	302.2	Komunikasi	55
17	800	Kesusastraan	22
18	700	Kesenian dan Rekreasi	7
		Total	2742

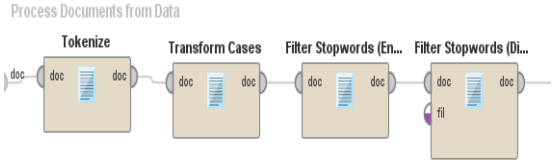
3.4 Data Mining

Untuk memprosesan data mining peneliti menggunakan software Rapidminer Studio versi 9.7.



Gambar 3. Pengolahan Data dengan Rapid Miner

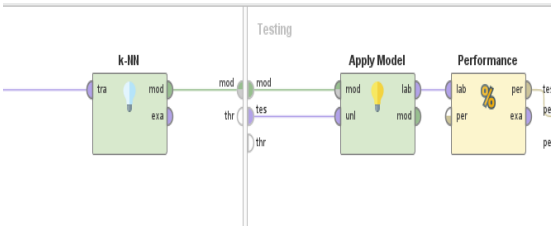
Sebelum pengolahan data ada 5 buah operator yang dibutuhkan yaitu, retrieve data buku, Set Role, Nominal to Text, Proses Document, Cross Validation. Operator retrieve data buku digunakan untuk membaca data buku yang akan di proses, operator Set Role digunakan untuk menentukan field atau bagian yang akan digunakan sebagai label. Pada penelitian ini atribut yang digunakan adalah atribut text dan atribut label. Set role juga digunakan untuk mengubah atribut role (misal, regular, special, label, id, dll). Operator nominal to text digunakan untuk mengubah jenis atribut nominal yang dipilih untuk teks. Serta memetakan semua nilai atribut ke dalam nilai string yang sesuai. Operator Process Documen From Data digunakan untuk membersihkan data agar menjadi vector, vector yang digunakan adalah TF-IDF, yang dapat digunakan untuk perhitungan algoritma diantaranya adalah seperti subproses dari process dokumen form data gambar di bawah berikut :



Gambar 4. Sub proses dari Operator Proses Document From Data

Di dalam operator proses document terdapat operator Tokenize, Transform Cases, Filter

Stopword English, Filter Stopword Dictionary. Setelah dokumen menjadi vector yang dapat dihitung karena nilai teks terebut menjadi nominal maka selanjutnya adalah proses validasi dengan cross validation, k-fold yang digunakan adalah K Fold=10. Di dalam operator cross validation terdapat subproses yang berisi operator validasi menggunakan algoritma K-NN dan ukuran kedekatan yang digunakan adalah cosine similarty, dan operator apply model yang digunakan untuk membuat model dari proses validasi, serta operator performance yang digunakan untuk mengukur kinerja dari model yang terbentuk.



Gambar 5. Sub proses dari Operator Cross Validation

Di dalam operator Cross validation terdapat operator KNN, Apply Model, Performance.

Berikut ini adalah hasil penggujian akurasi dengan menggunakan nilai k yang berbeda :

k	1	2	3	4	5	6	7	8	9
Accuracy (%)	60.83	60.79	66.59	68.53	69.18	69.62	70.28	71.23	72.07
k	10	11	12	13	14	15	16	17	18
Accuracy (%)	72.17	72.28	72.50	72.39	72.39	72.43	72.39	72.32	72.43

Gambar 6. Hasil Pengujian K=1 sampai K=18

Dari hasil pengujian seperti pada gambar 6 di atas dapat dilihat bahwa nilai akurasi semakin meningkat dengan bertambahnya nilai k hal ini di sebabkan karena dimensi klasifikasi pada label dalam penelitian ini ada sebanyak 18 klasifikasi. Pada kasus penelitian ini dengan menambahkan nilai k maka semakin banyak pula rujukan yang dapat di gunakan untuk menentukan sebuah dokumen termasuk ke dalam kelompok kelas yang mana dalam suatu label atribut klasifikasi. Pada Tabel di atas dapat dilihat bahwa nilai akurasi tertinggi terdapat di k = 12 yaitu sebesar 72.50%.

3.5 Interpretation

Selain nilai akurasi pada penelitian ini juga dilihat nilai ukuran lain seperti precision dan recall, di karenakan distribusi data class yang digunakan tidak merata (imbalance) (seperti pada table 1). ukuran akurasi bekerja paling baik ketika kelas data di distribusikan secara merata. Pengukuran lain seperti sensitivity (atau recall), specificity, precision, F dan F_β lebih cocok untuk masalah keseimbangan kelas [5].

Class precision : dapat dianggap sebagai ukuran ketepatan (yaitu berapa persentase tuple yang diberi label positif sebenarnya).

Misal seperti pada tabel 2 Untuk kelas 2x0 class precision adalah 76.65% artinya sebanyak 805 tuple terprediksi ke dalam label class 2x0, tetapi hanya sebanyak 617 tuple yang masuk ke dalam label class 2x0 yang sebenarnya.

Dan class recall : adalah ukuran kelengkapan (berapa persentase tuple positif yang diberi label seperti itu)

Misal seperti pada tabel untuk kelas 2x0 class recallnya adalah 81.72% artinya dari 755 jumlah tuple 2x0, yang terprediksi benar adalah sebanyak 617 tuple.

Nilai dari precision dan recall biasanya digunakan bersama, dimana nilai precision dibandingkan dengan nilai recall, atau sebaliknya. Cara alternatifnya adalah dengan mengkombinasikan nilai precision dan recall menjadi satu ukuran dengan F measure (atau di kenal dengan F1 Score). F1 Score adalah harmonic mean dari nilai precision dan recall. Berikut adalah hasil dari F1 Score.

Tabel 3. Nilai F1 Score k=12

class	precision	recall	F1 Score
2x0	76.65%	81.72%	79.10%
070	100.00%	44.44%	61.53%
370	59.01%	70.11%	64.08%
330	72.73%	63.31%	67.69%
300	61.11%	72.81%	66.45%
150	87.50%	68.14%	76.62%
2x2	76.12%	71.83%	73.91%
600	85.37%	53.03%	65.42%
400	92.00%	80.23%	85.71%
020	90.00%	72.97%	80.60%
2x1	88.98%	73.43%	80.46%
657	100.00%	86.96%	93.03%
500	93.75%	63.83%	75.95%
340	67.20%	68.31%	67.75%
375	100.00%	60.00%	75.00%
302.2	86.96%	72.73%	79.21%
800	84.21%	72.73%	78.05%
700	57.14%	57.14%	57.14%

Dari table 3 di atas dapat dilihat bahwa dari nilai F1 Score, class yang mempunyai nilai > 72.50% akurasi, adalah 2x0, 150, 2x2, 400, 020, 2x1, 657, 500, 375, 302.2 800. Dan class yang mempunyai F1 Score < 72.50% akurasi, adalah 070, 370, 330, 300, 600, 340, 700.

4. Kesimpulan

Berdasarkan hasil analisis data dan pembahasan, dapat diuraikan kesimpulan yang di dapat dari hasil penelitian ini adalah :

1. Untuk memberikan tingkat kesalahan minimum nilai k pada Nearest Neigbor Classifier dapat dipilih dengan cara melakukan serangkaian set pengujian pada dataset.
2. Semakin besar dimensi klasifikasi pada label kelas maka nilai k akan semakin bertambah, dengan menambahkan nilai k maka semakin banyak pula rujukan yang dapat digunakan untuk menentukan sebuah dokumen masuk ke dalam suatu kelas.
3. Untuk dataset yang memiliki data kelas yang berdistribusi tidak merata (imbalance) maka evaluasi model klasifikasi tidak hanya dilihat dari nilai akurasi, nilai recall atau sensitivity, nilai precision, dan F measure.
4. Nilai k = 12 memiliki akurasi sebesar 72.50%, jika dilihat dari nilai F1 Score, maka kelas terklasifikasi yang nilainya lebih dari 70% adalah 2x0, 150, 2x2, 400, 020, 2x1, 657, 500, 375, 302.2 800, dan dapat digunakan untuk mengklasifikasikan buku dengan label klasifikasi tersebut. Sebaliknya untuk label kelas 070, 370, 330, 300, 600, 340, 700, tidak dapat di jadikan rujukan untuk klasifikasi buku, karena memiliki nilai F1 score dibawah 70%.

Daftar Pustaka

[1] F. A. Wiranto, Supriyanto, dan Suryaningsih, “Perpustakaan Menjawab Tantangan Jaman,” 1997.

[2] M. R. Herga, “IMPLEMENTASI TEXT MINING SISTEM KLASIFIKASI DAN PENCARIAN KONTEN BUKU PERPUSTAKAAN MENGGUNAKAN ALGORITMA,” hlm. 6.

[3] E. K. Putri dan T. Setiadi, “PENERAPAN TEXT MINING PADA SISTEM KLASIFIKASI,” vol. 2, hlm. 11, 2014.

[4] D. T. Larose, *Discovering Knowledge in Data : an Introduction to Data Mining*. Canada: John Wiley & Sons, Inc., 2005.

[5] J. Han, J. Pei, dan M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] G. Gan, C. Ma, dan J. Wu, *Data clustering: theory, algorithms, and applications*. SIAM, 2007.

[7] R. Feldman dan J. Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[8] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, dan D. Delen, *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press, 2012.