p-ISSN: 2301-5373
e-ISSN: 2654-5101

Klasifikasi Jurnal menggunakan Metode KNN dengan Mengimplementasikan Perbandingan Seleksi Fitur

Farin Istighfarizky[a1], Ngurah Agus Sanjaya ER[a2], I Made Widiartha[a3], Luh Gede Astuti[a4], I Gusti Ngurah Anom Cahyadi Putra[a5], I Ketut Gede Suhartana[a6]

[a]Program Studi Informatika, Fakultas Matematika dan Ilmu Pengetahuan Alam,
Universitas Udayana
Bali, Indonesia
[1]farin79.istighfarizky@gmail.com
[2]agus_sanjaya@unud.ac.id
[3]madewidiartha@unud.ac.id
[4] lg.astuti@unud.ac.id
[5] anom.cp@unud.ac.id
[6]ikg.suhartana@unud.ac.id

Abstract

Classification is a process that automatically places text documents into a text based on the content of the text. Classification can help us classifying many text documents that have been published, with the classification, these text documents can be reached easily and quickly. Feature selection can be used to improve the performance of text classification in terms of learning speed and effectiveness. In the Chi-Square feature selection experiment, a 1% threshold combination with a parameter value of k=6 is the combination chosen to be the best model. In testing the new data, the K-Nearest Neighbor model by selecting the Chi-Square feature produces precision performance, recall, F1-Score, and accuracy respectively, namely 85%, 83.3%, 88.2%, and 92.3%. In the Gini Index feature selection experiment,1% threshold combination with a parameter value of k=4 is the combination chosen to be the best model. This threshold selects about 31 features with the highest Gini Index value. In testing the new data, the K-Nearest Neighbor model by selecting the Gini Index feature produces precision performance, recall, F1-Score, and accuracy respectively, namely 81.2%, 80.3%, 81.6%, and 86.6%.

Keywords: *Classification, Chi-Square, Gini Index, Features Selection*, *K-Nearest Neighbor*

1. Pendahuluan

Penemuan dibidang informatika mengalami perkembangan yang sangat pesat. Ratusan penelitian dilakukan di berbagai bidang disetiap tahunnya yang mana dengan harapan hasil penelitian tersebut dapat digunakan untuk penemuan berikutnya. Tidak semua penemuan akan relevan terhadap penelitian yang dilakukan oleh seseorang, oleh karena itu diperlukannya pengelompokan penelitian agar lebih mudah dalam mencari penelitian yang kita inginkan atau butuhkan. Jumlah studi yang dipublikasikan ada ratusan bahkan ribuan penelitian setiap tahunnya, seseorang akan membutuhkan terlalu banyak usaha dan dana yang besar dalam mengelompokkan jurnal-jurnal penelitian. Masalah ini dapat diselesaikan dengan klasifikasi teks.

Klasifikasi adalah proses menempatkan dokumen teks secara otomatis ke dalam kategori berdasarkan teks [1]. Klasifikasi membantu mengklasifikasikan jumlah dokumen teks yang diterbitkan. Klasifikasi membuatnya cepat dan mudah untuk mengelompokkan dokumen tekstual. Dasar dari algoritma seleksi fitur adalah untuk menemukan semua kemungkinan kombinasi atribut dalam data. Ini digunakan untuk menemukan subset terbaik untuk prediksi. Pemilihan fitur dapat digunakan untuk meningkatkan kinerja klasifikasi teks dalam hal kecepatan dan efektivitas pembelajaran [2].

Studi klasifikasi telah dilakukan oleh beberapa peneliti, seperti tentang mengkategorikan soal ujian secara otomatis. Pada penelitian tersebut, penulis menggunakan metode KNN dengan seleksi fitur *Chi-Square*. Hasil pada penelitian tersebut menunjukkan bahwa metode seleksi fitur *Chi-Square* terbukti mampu meningkatkan performa dari metode KNN [3]. Penelitian selanjutnya adalah

membahas tentang kognitif soal pada taksonomi *bloom* dengan KNN. Hasil yang didapatkan dari algoritma KNN dengan seleksi fitur *Gini Index* pada penelitian tersebut adalah akurasi sebesar 68,37% dan kappa tertinggi sebesar 0,607. Berdasarkan hasil tersebut, *Gini Index* mampu mengurangi dimensi fitur yang tinggi [4].

Seleksi fitur *Chi Square* menggunakan teori statistik untuk menentukan independensi suatu *term* dari kategorinya. Dalam seleksi fitur *Chi Square* berdasarkan teori statistika, dua peristiwa di antaranya adalah kemunculan dari fitur dan kemunculan dari kategori yang kemudian nilai *term* diurutkan dari yang tertinggi. Sedangkan seleksi fitur *Gini Index* mampu mengurangi dimensi fitur yang tinggi pada klasifikasi teks.

Berdasarkan penelitian yang dilakukan sebelumnya, pada penelitian kali ini penulis melakukan klasifikasi jurnal menggunakan metode *K-Nearest Neighbor* dengan menggunakan dua seleksi fitur yaitu *Chi-Square* dan *Gini Index*. Penulis berharap bahwa dengan menggunakan kombinasi metode ini dapat menghasilkan performa *precision*, *recall*, *f1-score*, dan akurasi yang lebih baik dibandingkan penelitian sebelumnya.

2. Metode Penelitian

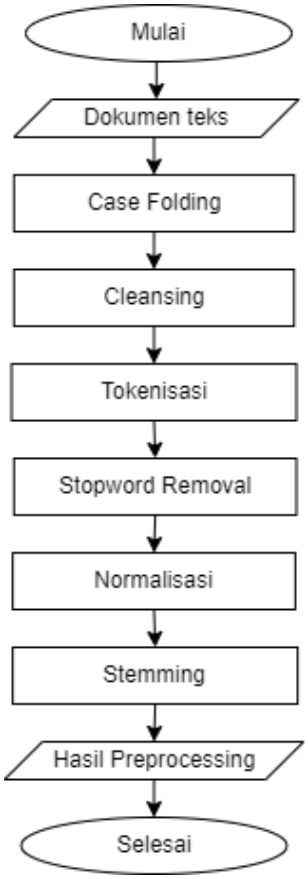
2.2. Dataset

Data yang digunakan pada penelitian ini adalah jurnal yang dipublikasikan oleh SINTA (*Science and Technology Index*) dan *Google Scholar*, data dapat diperoleh dari *website* SINTA (*Science and Technology*) (https://sinta.ristekbrin.go.id/) dan *Google Scholar* (https://scholar.google.com) Kemudian data yang digunakan untuk proses klasifikasi adalah pada bagian teks abstrak jurnal yang berbahasa Indonesia yang disimpan dalam bentuk (.xlsx) untuk digunakan sebagai data latih dan data uji. Dokumen yang digunakan berjumlah 100 data per kelas disetiap artikel ilmiah yaitu: pendidikan, ekonomi, dan informatika. Data *testing* yang diuji sebanyak 60 data jurnal dan data *training* sebanyak 240 data jurnal. Data pada penelitian ini adalah data sekunder.

2.2. *Preprocessing*

Preprocessing proses pertama mempersiapkan dataset sebelum pembobotan, tujuannya adalah untuk menyederhanakan pemrosesan data dan juga untuk mendapatkan tingkat performa yang tinggi. Proses *preprocessing* dapat dilihat pada gambar 1.

p-ISSN: 2301-5373
e-ISSN: 2654-5101

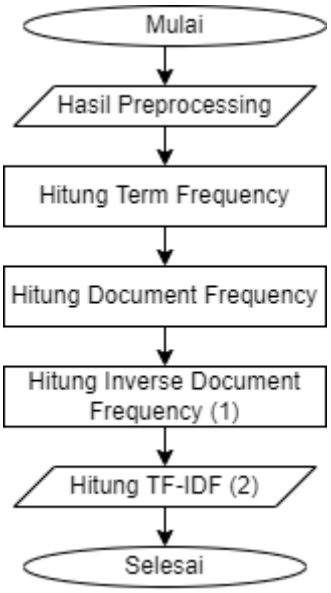


Gambar 1. Proses *preprocessing*

Pada tahap ini, terdapat beberapa proses yaitu *case folding, cleansing,* tokenisasi*, stopword,* normalisasi, dan *stemming*. *Case folding* yaitu proses mengubah semua huruf menjadi huruf kecil. *Cleansing* merupakan proses penghapusan karakter yang tidak relevan dengan klasifikasi jurnal. Tokenisasi yaitu pemisahan kata-kata paragraf atau kalimat menjadi token-token tertentu. *Stopword removal* yaitu penghapusan kata yang tidak mempengaruhi klasifikasi jurnal. Proses normalisasi yaitu mengubah dan mengembalikan bentuk penulisan tidak baku ke bentuk penulisan yang sesuai dengan KBBI. Proses terakhir adalah *stemming,* yaitu mengekstrak kata yang dilampirkan ke dalam kata dasar [5].

2.3. Term Frequency Inverse-Document Frequency (TF-IDF)

Metode *Term Frequency Invers Document Frequency* adalah metode pembobotan yang menggabungkan frekuensi istilah dalam satu set dokumen dan kelangkaannya. Metode ini menggabungkan dua konsep pembobotan, yaitu frekuensi kemunculan kata dalam dokumen tertentu dan frekuensi kebalikan dari dokumen yang berisi kata tersebut. Berapa kali sebuah kata muncul dalam dokumen tertentu menunjukkan pentingnya kata tersebut dalam dokumen itu. Frekuensi dokumen yang berisi kata-kata menunjukkan seberapa sering kata-kata itu muncul. Sehingga bobot hubungan antara sebuah kata dan sebuah dokumen akan tinggi apabila frekuensi kata tersebut tinggi di dalam dokumen dan frekuensi keseluruhan dokumen yang mengandung kata tersebut yang rendah pada kumpulan dokumen. Proses TF-IDF dapat dilihat pada gambar 2.



Gambar 2. TF-IDF

a. Hitung jumlah kemunculan *term* i dalam dokumen j (tfi,j).
b. Hitung jumlah dokumen yang mengandung *term* i (df)
c. Menghitung nilai bobot *inverse document frequency* (idf) dengan menggunakan persamaan:

$$idf_i = log\left(\frac{N}{df_i}\right) \quad (1)$$

Keterangan:
N = jumlah dokumen secara keseluruhan

d. Menghitung nilai bobot TF-IDF dengan menggunakan persamaan:

$$w_{i,j} = tf_{i,j}\ x\ idf_i \quad (2)$$

Keterangan:
$w_{i,j}$ = bobot *term* i terhadap dokumen j
$tf_{i,j}$ = frekuensi *term* i pada dokumen j
idf_i = nilai bobot IDF pada *term* i

2.4. *K-Nearest Neighbor* (KNN)

Algoritma *K-Nearest Neighbor* adalah metode untuk mengklasifikasi objek berdasarkan data latih yang paling dekat dengan objek tersebut. Metode *K-Nearest Neighbor* adalah algoritma pembelajaran terawasi, dan hasil dari *query instance* baru dikategorikan berdasarkan sebagaian besar kategori algoritma *K-Nearest Neighbor*. Kelas yang paling sering ditampilkan adalah kelas yang diperoleh dari hasil klasifikasi. Kedekatan didefinisikan dalam jarak metrik, seperti jarak *Euclidean* [6].

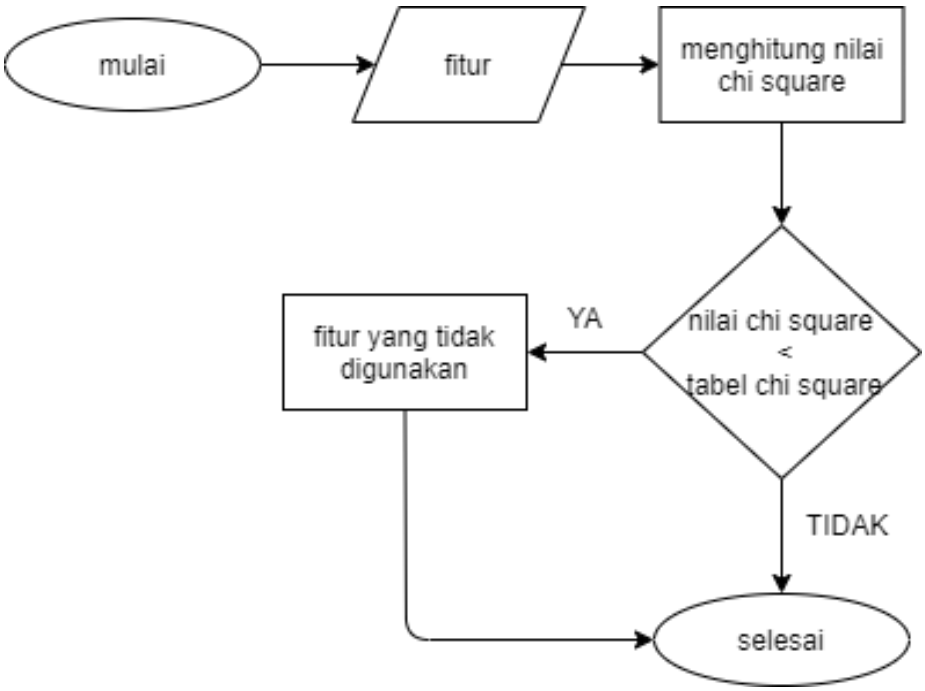
$$D_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

Keterangan:
D = jarak kedekatan
x = data training
y = data testing
n = jumlah atribut individu antara 1 s.d n
i = atribut individu antara 1 s.d n

p-ISSN: 2301-5373
e-ISSN: 2654-5101

2.5. Chi-Square

Chi-Square adalah metode untuk menghitung ketergantungan fitur. Pada pemrosesan teks biasanya menggunakan dua kelas untuk mengukur ketergantungan antara dua label dan kata – antara kelas tertentu c.Tahapan *Chi-Square* dapat dilihat pada gambar 3.



Gambar 3. *Chi-Square*

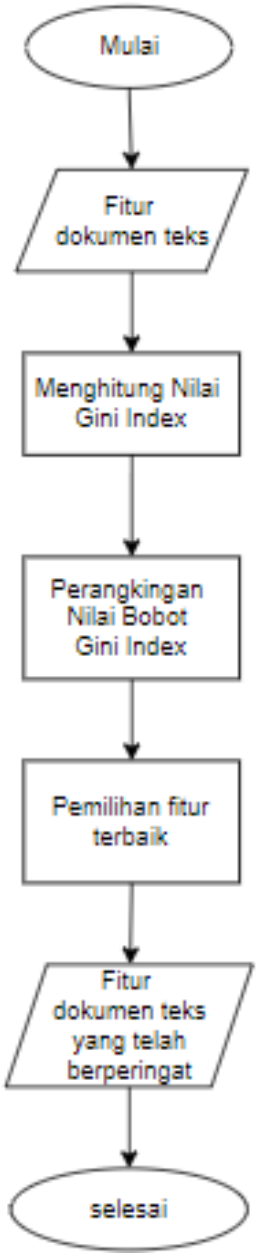
Berikut ini adalah rumus yang digunakan untuk menerapkan metode seleksi fitur *Chi-Square* [7].

$$X^2(t,c) = \frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (4)$$

Keterangan:
t = kata
c = kelas/kategori
N = jumlah data latih
A = jumlah dokumen pada kelas c yang memuat t,
B = jumlah dokumen yang tidak ditemukan pada kelas c tapi memuat t,
C = jumlah dokumen pada kelas c yang tidak memuat t,
D = jumlah dokumen yang bukan merupakan dokumen kelas c dan tidak memuat *term* t

2.6. Gini Index

Gini Index adalah kriteria berbasis ketidakmurnian data yang mengukur perbedaan antara distribusi probabilitas dari nilai atribut. *Gini Index* umumnya dipakai dalam Algoritma *Classification and Regression Trees* yang merepsentasikan ukuran seberapa acak pilihan objek dari data latih. Ukuran ketidak murnian mencapai 0 ketika hanya 1 kelas saja yang ada pada sebuah titik. Namun sebaliknya akan mencapai maksimum ketika ukuran kelas pada titik tersebut seimbang. *Gini Index* dapat dianggap sebagai probabilitas dari dua data yang dipilih secara acak dari kelas yang berbeda akan digunakan dalam penelitian ini untuk mengukur divergensi yang digunakan sebagai dasar bobot setiap. Cocok untuk pemilahan, sistem biner, nilai numerik terus menerus, dan lain-lain. Tahapan *Gini Index* dapat dilihat pada gambar 4.



Gambar 4. *Gini Index*

Berikut ini adalah rumus yang digunakan untuk menerapkan metode seleksi fitur *Gini Index* [8].

$$GI(t) = 1 - \sum_{i=1}^{C}[p(i|t)]^2 \quad (5)$$

Keterangan:
C = total kelas
t = *term*
p(i|t) = peluang kelas i terhadap term t

2.7. Evaluasi

Confusion matrix merupakansalah satu metode yang dapat digunakanuntuk mengukur performansi suatu metode klasifikasi. Pada dasarnya *confusion matrix* ini berisi informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi sebagaimana mestinya. Saat mengukur kinerja menggunakan *confusion matrix*, ada empat istilah yang menggambarkan hasil dari proses klasifikasi. Keempat istilah tersebut adalah *True Positive, True Negative, False Positive*,

p-ISSN: 2301-5373
e-ISSN: 2654-5101

dan *False Negative*. Nilai *True Negative* (TN) adalah jumlah data negatif yang terdeteksi dengan benar, dan *False Positive* (FP) adalah data negatif tetapi terdeteksi sebagai data positif. Sementara itu, *True Positive* (TP) di sisi lain adalah data positif yang dikenali dengan benar. *False Negative* (FN) adalah kebalikan dari *True Positive*, sehingga datanya positif tetapi dikenali sebagai data negatif. Tabel *Confusion matrix* dapat dilihat pada tabel 1 [9].

Tabel 1. *Confusion Matrix*

Kelas	Terklarifikasi Positif	Terklarifikasi Negatif
Positif	TP (*True Positive*)	FN (*False Negative*)
Negatif	FP (*False Positive*)	TN (*True Negative*)

Dimana:

- *True Positive*, jumlah data positif yang diklasifikasikan dengan benar oleh sistem.
- *True Negative*, jumlah data negatif yang diklasifikasikan dengan benar oleh sistem.
- *False Negative*, jumlah data negatif namun diklasifikasikan salah oleh sistem.
- *False Positive*, jumlah data positif namun diklasifikasikan salah oleh sistem.

Berdasarkan nilai *True Negative* (TN), *False Positive* (FP*), False Negative* (FN), dan *True Positive* (TP) dapat diperolehnilai akurasi, presisi dan*recall*. Nilai akurasi menggambarkan seberapa akurat sistem dapat mengklasifikasikan data secara benar.Dengan kata lain,nilai akurasi merupa kan perbandingan antara data yang terklasifikasi benar dengan keseluruhan data. Nilai akurasi dapat diperoleh dengan persamaan (6). Nilai presisi menggambarkan jumlah data kategori positif yang diklasifikasikan secara benar dibagi dengan total data yang diklasifikasi positif. Presisi dapat diperoleh dengan persamaan (7) sementara itu, *recall* menunjukkan berapa persen data kategori positif yang terklasifikasikan dengan benar oleh sistem. Nilai *recall* diperoleh dengan persamaan 8.

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} * 100\% \quad (6)$$

$$\text{Presisi} = \frac{TP}{FP+TP} * 100\% \quad (7)$$

$$\textit{Recall} = \frac{TP}{FN+TP} * 100\% \quad (8)$$

Setelah mendapat nilai *recal*l dan *precision*, maka dilakukan perhitungan menggunakan F1-*score*. F1-*score* digunakan untuk mengukur kombinasi hasil *precision* dan *recall*, sehingga menjadi satu nilai pengukuran. F1-*Score* dapat dihitung menggunakan persamaan 9:

$$F1 = \frac{2}{\frac{1}{Precission} + \frac{1}{Recall}} \quad (9)$$

3. Hasil dan Pembahasan

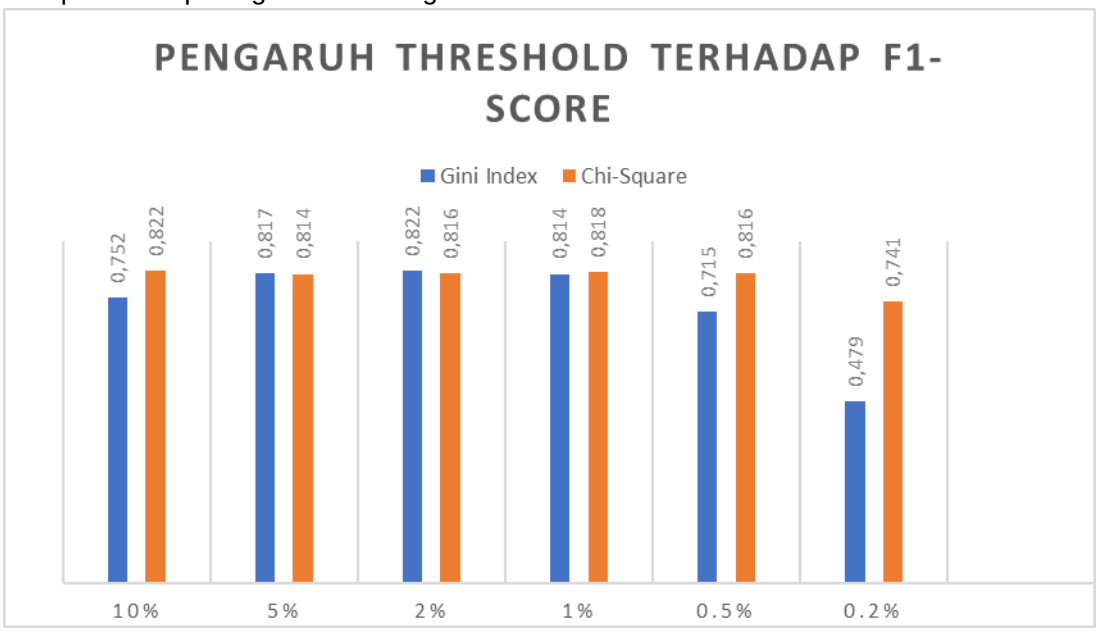
Pada penelitian yang dilakukan ada 80% total data digunakan selama tahap pelatihan dan juga validasi. Perubahan nilai k pada percobaan adalah k = 4, k = 6, k = 7, k = 9, dan k = 11. Uji *threshold* dilakukan dengan menggunakan seleksi fitur *Gini Index* dan seleksi fitur *Chi-Square*. *Threshold* adalah persentase jumlah fitur yang dipilih dari semua fitur yang diurutkan. *Threshold* yang digunakan adalah 10%, 5%, 2%, 1%, 0.5%, dan 0.2%. Pada setiap iterasi dari 10-*Fold Cross Validation*, akan dihitung rata-rata performa F1-*Score* dan akurasi dengan menggunakan persamaan (9) dan (6). Nilai k dengan kinerja F1-*Score* tertinggi akan dipilih sebagai model yang terbaik. Nilai k dengan F1-*Score* tertinggi berarti hasil klasifikasi jurnal lebih akurat. Setelah melakukan proses pelatihan dan validasi pada model *K-Nearest Neighbor* menggunakan uji 10-*Fold Cross Validation*,

didapatkan nilai k dengan performansi F1-*Score* terbaik. Uji kombinasi *threshold Chi-Square* dan *Gini Index* dan nilai k yang sudah dilakukan menghasilkan beragam performa yang berbeda. Dari hasil pengujian yang diperoleh, dapat dilihat pengaruh dari *threshold* yang digunakan untuk performansi dari metode *K-Nearest Neighbor*, seperti yang ditunjukkan pada tabel 5.

Tabel 5. Hasil Evaluasi Pengujian KNN dengan Seleksi Fitur Chi-Square dan Gini Index

Threshold	**Ukuran Evaluasi (Rata-Rata Fold)**			
	F-1 Score Chi-Square	**F1-Sscore Gini Index**	**Akurasi Chi-Square**	**Akurasi Gini Index**
10%	82,2%	75,2%	82%	75%
5%	81,4%	81,7%	81,2%	81,6%
2%	81,6%	82,2%	81,2%	82%
1%	81,8%	81,4%	81,6%	80,8%
0.5%	81,6%	71,5%	81,2%	71,6%
0.2%	74,1%	47,9%	74,1%	48,3%

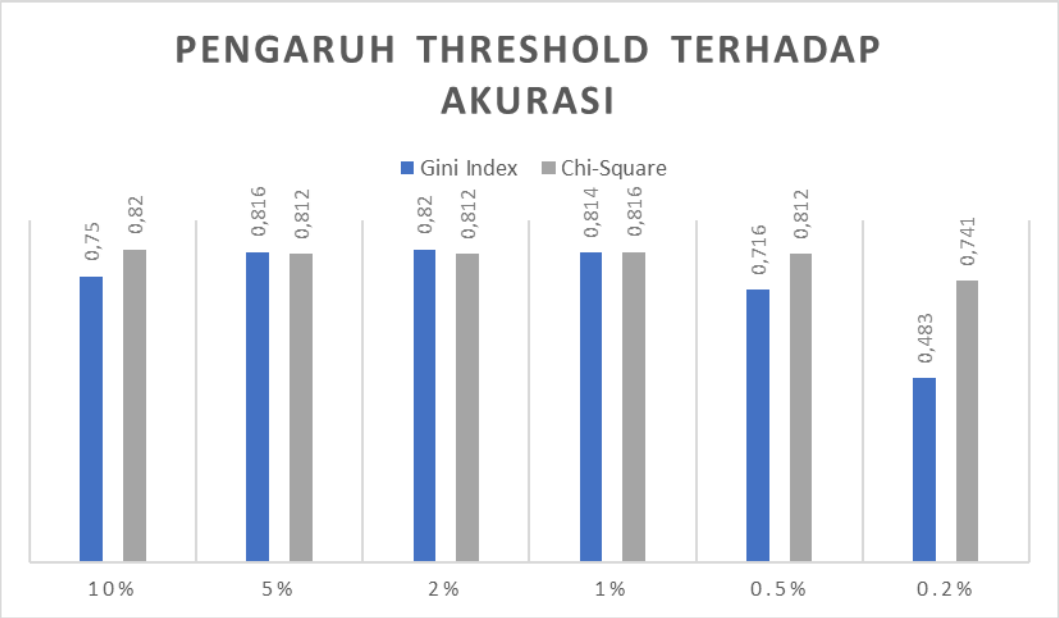
Pengujian kombinasi *threshold Chi-Square* dan *Gini Index* dan nilai k yang sudah dilakukan menghasilkan beragam performa yang berbeda. Dari hasil pengujian yang diperoleh dapat diketahui pengaruh *threshold* yang digunakan terhadap evaluasi kinerja metode *K-Nearest Neighbor*, hal tersebut dapat dilihat pada gambar 5 dan gambar 6.



Gambar 5. Pengaruh Threshold terhadap F1-Score

Gambar 5 menunjukkan pengaruh *threshold* terhadap performa *F1-Score* metode *K-Nearest Neighbor* dengan seleksi fitur *Chi-Square* dan seleksi fitur *Gini Index*. *F1-Score* yang tertera pada Gambar 5 adalah nilai *F1-Score* dari kombinasi nilai k dengan akurasi tertinggi. Nilai *threshold* diketahui memiliki pengaruh terhadap *F1-Score*.

p-ISSN: 2301-5373
e-ISSN: 2654-5101



Gambar 6. Pengaruh Threshold terhadap akurasi

Akurasi yang tertera pada Gambar 6 adalah akurasi tertinggi dari kombinasi nilai k yang sudah diuji. Nilai threshold diketahui memiliki pengaruh terhadap akurasi. Model dengan threshold yang menghasilkan akurasi di atas nilai rata-rata tersebut dapat dikatakan sebagai model dengan performa yang baik dalam mengklasifikasikan jurnal.

4. Kesimpulan

Berdasarkan penelitian yang telah dilakukan, ditarik kesimpulan bahwa implementasi seleksi fitur dapat meningkatkan performa *precision*, *recall, F1-Score,* dan akurasi dari metode *K-Nearest Neighbor* dalam mengklasifikasikan jurnal dan seleksi fitur *Chi-Square* lebih unggul daripada seleksi fitur *Gini Index*. Pada eksperimen seleksi fitur *Chi-Square*, kombinasi *threshold* 1% dengan parameter nilai k=6 adalah kombinasi yang dipilih menjadi model terbaik. *Threshold* ini menyeleksi sekitar 31 fitur dengan nilai *Chi-Square* tertinggi. Pada pengujian data baru, model *K-Nearest Neighbor* dengan seleksi fitur *Chi-Square* menghasilkan performa *precision*, *recall, F1-Score,* dan akurasi secara berturut-turut yaitu 85%, 83.3%, 88.2%, dan 92.3%. Pada eksperimen seleksi fitur *Gini Index*, kombinasi *threshold 1*% dengan parameter nilai k=4 adalah kombinasi yang dipilih menjadi model terbaik. *Threshold* ini menyeleksi sekitar 31 fitur dengan nilai *Gini Index* tertinggi. Pada pengujian data baru, model *K-Nearest Neighbor* dengan seleksi fitur *Gini Index* menghasilkan performa *precision*, *recall, F1-Score,* dan akurasi secara berturut-turut yaitu 81.2%, 80.3%, 81.6%, dan 86.6%

Daftar Pustaka

[1] Z. XIONG, J. JIANG and Y. ZHANG, "New feature selection approach (CDF) for text categorization", *Journal of Computer Applications*, vol. 29, no. 7, pp. 1755-1757, 2009. Available: 10.3724/sp.j.1087.2009.01755.

[2] H. Alshalabi, S. Tiun, N. Omar and M. Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", *Procedia Technology*, vol. 11, pp. 748-754, 2013. Available: 10.1016/j.protcy.2013.12.254.

[3] I. Listiowarni and N. Puspa Dewi, "Pemanfaatan Klasifikasi Soal Biologi Cognitive Domain Bloom's Taxonomy Menggunakan KNN Chi-Square Sebagai Penyusunan Naskah Soal", *Digital Zone: Jurnal Teknologi Informasi dan Komunikasi*, vol. 11, no. 2, pp. 186-197, 2020. Available:

10.31849/digitalzone.v11i2.4798.

[4] T. Setiyorini and R. Asmono, "PENERAPAN METODE K-NEAREST NEIGHBOR DAN GINI INDEX PADA KLASIFIKASI KINERJA SISWA", *Jurnal Techno Nusa Mandiri*, vol. 16, no. 2, pp. 121-126, 2019. Available: 10.33480/techno.v16i2.747.

[5] K. Yonatha Wijaya and A. Karyawati, "The Effects of Different Kernels in SVM Sentiment Analysis on Mass Social Distancing", *JELIKU (Jurnal Elektronik Ilmu Komputer Udayana)*, vol. 9, no. 2, p. 161, 2020. Available: 10.24843/jlk.2020.v09.i02.p01.

[6] H. Hadi and T. Sukamto, "Klasifikasi Jenis Laporan Masyarakat Dengan K-Nearest Neighbor Algorithm", *JOINS (Journal of Information System)*, vol. 5, no. 1, pp. 77-85, 2020. Available: 10.33633/joins.v5i1.3355.

[7] C. Suharno, M. Fauzi and R. Perdana, "Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-square", *Systemic: Information System and Informatics Journal*, vol. 3, no. 1, pp. 25-32, 2017. Available: 10.29080/systemic.v3i1.191.

[8] C. Aggarwal, Data Mining: The Textbook. Springer International Publishing Switzerland, 2015. Available: 10.1007/978-3-319-14142-8.

[9] M. Imron and B. Prasetyo, "Improving Algorithm Accuracy K-Nearest Neighbor Using Z-Score Normalization and Particle Swarm Optimization to Predict Customer Churn", *Shmpublisher.com*, 2022. [Online]. Available: https://shmpublisher.com/index.php/joscex/article/view/7.