Jurnal Algoritma, Logika dan Komputasi *Vol.VII (No. 01) : 643 - 651. Th. 2024*
Versi Online: https://journal.ubm.ac.id/index.php/alu *p-ISSN: 2620-620X*
DOI: http://dx.doi.org/10.30813/j-alu.v2i2.6028 *e-ISSN: 2621-9840*

MEMPREDIKSI PENINGKATAN H-INDEKS UNTUK JURNAL PENELITIAN DENGAN MENGGUNAKAN ALGORITMA COST-SENSITIVE SELECTIVE NAIVE BAYES CLASSIFIERS

Predicting Increased H-Index For Research Journals Using The Cost-Sensitive Selective Naive Bayes Classifiers Algorithm

Reycardo Henglie, Reycardo.henglie@gmail.com[1)], Yunianto Purnomo, ypurnomo@bundamulia.ac.id[2)*] , Jusia Amanda Ginting, jginting@bundamulia.ac.id
[1)2)3)]Program Studi Informatika, Universitas Bunda Mulia, Jakarta Utara

Diterima 26 Juli 2024 / Disetujui 31 Juli 2024

Abstract

Machine learning community is not only interested in maximizing classification accuracy, but also in minimizing the distances between the actual and the predicted class. Some ideas, like the cost-sensitive learning approach, are proposed to face this problem. In this paper, we propose two greedy wrapper forward cost-sensitive selective naive Bayes approaches. Both approaches readjust the probability thresholds of each class to select the class with the minimum-expected cost. The first algorithm (CSSNB-Accuracy) considers adding each variable to the model and measures the performance of the resulting model on the training data. The variable that most improves the accuracy, that is, the percentage of well classified instances between the readjusted class and actual class, is permanently added to the model. In contrast, the second algorithm (CS-SNB-Cost) considers adding variables that reduce the misclassification cost, that is, the distance between the readjusted class and actual class. We have tested our algorithms on the bibliometric indices prediction area. Considering the popularity of the well-known h-index, we have researched and built several prediction models to forecast the annual increase of the h-index for Neurosciences journals in a four-year time horizon. Results show that our approaches, particularly CS-SNB-Accuracy, achieved higher accuracy values than the analyzed cost sensitive classifiers and Bayesian classifiers. Furthermore, we also noted that the CS-SNB-Cost always achieved a lower average cost than all analyzed cost-sensitive and cost-insensitive classifiers. These cost sensitive selective naive Bayes approaches outperform the selective naive Bayes in terms of accuracy and average cost, so the cost-sensitive learning approach could be also applied in different probabilistic classification approaches.

Keyword: *CSSNB-Accuracy, CS-SNB-Cost, bibliometric, clasification, predicted distances*

Abstrak

Komunitas *Machine learning* tidak hanya memaksimalkan akurasi dari klasifikasi, tetapi juga mengurangi rentang antara prediksi dan kenyataan. Ide seperti cost-sensitive learning digunakan untuk mengatasi masalah ini. Dalam jurnal ini, kami menggunakan dua algoritma cost-sensitive selective naive bayes. Algoritma pertama (CSSNB-Accuracy) dengan menambahkan setiap variabel ke dalam model dan mengukur hasil dari pelatihan data. Variabel yang memberikan akurasi terbaik, yaitu persentase dari contoh-contoh yang terklasifikasi antara kelas yang dibuat dengan kelas yang sebenarnya. Algoritma kedua (CS-SNB-COST) dengan menambahkan variabel yang hanya mengurangi cost kesalahan klasifikasi antara kelas yang dibuat dengan kelas yang sebenarnya. Penulis telah menguji coba algoritma pada prediksi indeks bibliometrik. Mengingat h-indeks yang telah banyak digunakan, penulis telah melakukan riset pada beberapa model prediksi untuk memprediksi peningkatan h-indeks untuk jurnal Neurosciences dalam empat tahun terakhir. Hasilnya algoritma CS-SNB-Accuracy, menghasilkan akurasi yang lebih tinggi dibandingkan cost-sensitive classifiers dan bayesian classifiers. Lalu cost-sensitive selective naive bayes menghasilkan performa yang lebih baik dibandingkan selective naive-bayes dalam hal akurasi dan cost rata-rata, sehingga cost sensitive learning dapat juga di gunakan dalam kemungkinan klasifikasi yang berbeda.

Kata kunci: *CSSNB-Accuracy, CS-SNB-Cost, bibliometric, klasifikasi, rentang prediksi*

*Korespondensi Penulis:
E-mail: **ypurnomo@bundamulia.ac.id**

PENDAHULUAN

Masalah dalam klasifikasi umumnya mengasumsi bahwa nilai-nilai kelas tidak berurutan. Namun, nilai-nilai ini memiliki tatanan alami dalam banyak aplikasi praktis. Mengingat kelas yang berurutan, penulis tidak hanya tertarik untuk memaksimalkan akurasi klasifikasi, tetapi juga dalam meminimalkan jarak antara kelas yang sebenarnya dan yang diprediksi. Beberapa bidang, seperti statistik, telah menghadapi masalah ini selama bertahun-tahun.

Penulis menggabungkan cost-sensitive learning dan fitur subset seleksi ke dalam naive bayes, yang merupakan metode yang paling mudah dan banyak diuji untuk induksi probabilistik dan telah lama digunakan dalam bidang pengenalan pola . Untuk alasan ini, penulis mengembangkan algoritma cost-sensitive baru berdasarkan pada gagasan elective naive Bayes. Khususnya, penulis mengembangkan dua algoritma langsung yang menambahkan biaya misklasifikasi ke algoritma pembelajaran, dan menggunakan pendekatan wrapper untuk memilih variabel yang relevan yang memaksimalkan akurasi (CS-SNB-Accuracy algorithm) dan meminimalkan biaya (CS-SNB-Cost algorithm). Tujuan dari pendekatan ini adalah untuk membangun model parsimonius. Model-model ini tidak akan menyertakan fitur-fitur yang tidak relevan dan berlebihan. Beberapa manfaat penerapan seleksi variabel adalah kinerja klasifikasi yang lebih baik, model klasifikasi yang lebih cepat, basis data yang lebih kecil, dan kemampuan untuk mendapatkan lebih banyak wawasan ke dalam proses yang dimodelkan.

Minat dan orisinalitas dari penelitian kami dibagi menjadi dua. Pertama, kami mengembangkan dua klasifikasi baru (CS-SNB-Accuracy dan CS-SNB-Cost) yang membawa keuntungan dalam menggunakan pendekatan cost-sensitive learning dan fitur pemilihan subset. Kedua, kedua klasifikasi digunakan untuk memprediksi peningkatan tahunan indeks-h untuk jurnal ilmiah yang termasuk dalam kategori Jurnal Sitasi Laporan Neurosciences di sepanjang waktu 4 tahun menggunakan indeks bibliometrik.

TINJAUAN TEORI

H-Index atau Indeks-h merupakan sebuah tolok ukur bagi seorang ilmuwan baik itu dosen ataupun peneliti dalam mengembangkan hasil karya keilmuwannya. Karya keilmuwan seorang dosen dan peneliti antara lain berupa hasil penelitian yang dipublikasikan, hak paten atau HKI (Hak Kekayaan Intelektual) dan artikel-artikel yang diseminarkan dalam bentuk jurnal ilmiah, baik Seminar Nasional maupun Internasional. Indeks-H pertama kali diperkenalkan oleh seorang fisikawan dari Universitas di California, San Diego, bernama Jorge E Hirsch pada tahun 1985 yang lalu ini dianggap sebagai cara yang paling efektif untuk menilai kinerja seorang ilmuwan pada saat ini. Namun perlu diketahui bahwa indeks-H ini masih terdapat kekurangan dalam validitasnya, seperti rentan terhadap manipulasi sitasi pribadi (self-citation).[1][2][3]

Indeks-H yang sering juga dikenal dengan Hirsch Index atau Hirsch Number ini dapat diperoleh di media pengindeks publikasi seperti: Portal Garuda, Google Scholar, DOAJ (Directory of Open Access Journals), EBSCO, CrossRef, BASE (Bielefeld Academic Search Engine), ISJD, ISJD, SINTA, Scopus, dsb.[4][5][6]

Algoritma Naive Bayes merupakan sebuah metode klasifikasi menggunakan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes. Algoritma Naive Bayes memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai Teorema Bayes. Ciri utama dari Naïve Bayes Classifier ini adalah asumsi yang sangat kuat (naïf) akan independensi dari masing-masing kondisi / kejadian.[6][7][8][9]

Naive Bayes Classifier bekerja sangat baik dibanding dengan model classifier lainnya. Naïve Bayes Classifier memiliki tingkat akurasi yang lebih baik dibanding model classifier lainnya".[10][11][12][13][14][15]

Keuntungan penggunan adalah bahwa metoda ini hanya membutuhkan jumlah data pelatihan (training data) yang kecil untuk menentukan estimasi parameter yang diperlukan dalam proses pengklasifikasian. Karena yang diasumsikan sebagai variabel independent, maka hanya varians dari suatu variabel dalam sebuah kelas yang dibutuhkan untuk menentukan klasifikasi, bukan keseluruhan dari matriks kovarians.[16][17][18]

Cost-sensitive Bayesian classifiers

Tujuan dari metode cost-sensitive adalah memperhitungkan biaya misklasifikasi yang berbeda dari 0 (hit) dan 1 (miss). Metode-metode ini berkaitan dengan akurasi klasifikasi dan biaya klasifikasi. Proses pencarian dari pendekatan pertama (CSSNB-Accuracy) didasarkan pada memaksimalkan akurasi klasifikasi, yaitu, termasuk variabel yang meningkatkan akurasi klasifikasi, sedangkan proses pencarian dari pendekatan kedua (CS-SNBCost) didasarkan pada meminimalkan biaya misklasifikasi, artinya, termasuk dalam variabel yang mengurangi jarak antara kelas yang sebenarnya dan yang diprediksi.[19][20][21][22][23][24] Dengan adanya matriks biaya dan sekumpulan probabilitas kelas yang diprediksi untuk setiap contoh, kedua pendekatan menyesuaikan kembali ambang kemungkinan setiap kelas untuk memilih kelas dengan biaya minimum yang diharapkan. Biaya yang diharapkan dari masing-masing prediksi diperoleh dengan mengalikan biaya yang terkait dengan probabilitas kelas yang diprediksi. Tidak seperti selective Naive Bayes, pendekatan ini tidak memilih nilai kelas yang paling mungkin dari distribusi posterior, mereka memilih kelas (cn) yang meminimalkan perkiraan biaya prediksi yang diberikan pada rumus (1) dan Rumus (2) [25][26][27][28][29]

$$c^* = \arg \min_{c \in \Omega(C)} \sum_{c' \in \Omega(C)} p(c'|\mathbf{x}) \quad cost(c|c') \tag{1}$$

Dimana

$$p(c'|\mathbf{x}) \propto p(c') \prod_{i=1}^{m} p(x_i|c') \prod_{j=m+1}^{n} \mathcal{N}(x_j, \mu_{c'j}, \sigma^2_{c'j}) \tag{2}$$

Cost(c|c') adalah biaya misklasifikasi terkait. Singkatnya, pendekatan pertama (CS-SNB-Accuracy) beranggapan dengan menambahkan setiap variabel ke model dan mengukur kinerja model yang dihasilkan pada data pelatihan. Variabel yang paling meningkatkan keakuratan, yaitu, persentase dari klasifikasi yang menghasilkan hasil yang baik dalam kelas prediksi dibandingkan kelas yang sebenarnya, secara permanen ditambahkan ke model. Sebaliknya, pendekatan kedua (CS-SNB-Cost) mempertimbangkan penambahan variabel yang mengurangi biaya misklasifikasi antara yang diprediksi dan kelas yang sebenarnya.[25][26][28]

Cost-sensitive selective naive Bayes – Accuracy(CS-SNB-Accuracy)

Algoritma ini memilih k-fold cross-validation sebagai prosedur untuk memperkirakan akurasi dan biaya model yang mengklasifikasikan kasus baru sesuai dengan nilai fitur prediktif. Metode ini di stratifikasi, yaitu membagi semua kasus menjadi himpunan bagian k dari proporsi yang kira-kira sama dari nilai kelas dan ukuran yang sama. Setiap bagian digunakan untuk menguji model yang dipelajari dari subset k-1 lainnya. [25][26][27][28][29]

Cost-sensitive selective naive Bayes – Cost(CS-SNB-Cost)

Algoritma ini juga memilih k-fold cross-validation sebagai prosedur untuk memperkirakan akurasi dan biaya model. Algoritma ini menginisialisasi model ke variabel kelas, yaitu, tidak ada variabel prediktif dalam model pada Tabel 1 [25][26][27][28][29]

Tabel 1. Korespondensi nilai h dan label kelas

Label Class	Tahun 1	Tahun 2	Tahun 3	Tahun 4
Low values	0-1	0-4	0-6	0-8
Medium-Low values	2	5-6	7-9	9-12
Medium-High Values	3-4	7-9	10-14	13-18
High Values	≥5	≥10	≥15	≥19

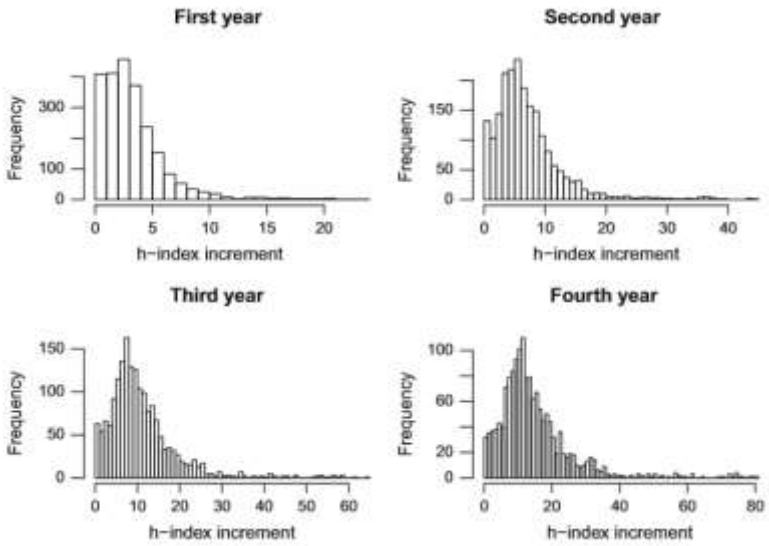
Metodologi Penelitian

Penulis telah memilih kategori Neurosciences untuk studi kasus Penulis. Penulis telah menggunakan platform Thomson Reuters 'Web of Science (WoS) dan Journal Citation Reports (JCR) untuk mengunduh publikasi dan data kutipan. Berikut ini, kami mengilustrasikan perbedaan tahapan dari konstruksi dataset dan menjelaskan setiap variabel dataset, yaitu fitur prediktif. [30][31][32][33][34][35][36]

Pertama, penulis memilih semua jurnal yang termasuk dalam JCR dengan kategori Neurosciences dari tahun 2000 hingga 2011. Ada 269 jurnal dalam kategori ini selama periode menganalisis. Kemudian penulis memperoleh daftar publikasi dan data kutipan untuk jurnal-jurnal ini dari WoS. Penulis mengunduh semua dokumen (1.044.811 makalah) yang diterbitkan oleh 269 jurnal hingga 2011. Menggunakan informasi di atas, penulis menghitung beberapa indeks dampak ilmiah yang terkait dengan jurnal yang dipilih untuk setiap jurnal dari tahun 2000 hingga 2011. penulis juga mengunduh nilai indeks jurnal spesifik lainnya dari JCR . Akhirnya, penulis menyimpan semua informasi dalam basis data yang dirancang untuk tujuan ini. [37][28][39]

HASIL DAN PEMBAHASAN

Penulis membandingkan algoritma yang telah penulis rancang dengan formulasi standar Naive Bayes tertentu untuk menentukan apakah akurasi dan nilai biaya rata-rata mereka masuk akal. Tabel 2 menunjukkan estimasi dari akurasi dan rata rata biaya untuk setiap model. Angka dalam huruf tebal mewakili nilai akurasi tertinggi dan nilai biaya terendah untuk setiap model.



Gambar 1. Distribusi Pertambahan h-Index

Penulis menguji metode penulis dengan matriks biaya yang berbeda (C(0,n), C(0,n^2), C(0,2^n), and C(0,n^n)). Harga matriks C(0,n) merepresentasikan harga-harga dimana hasil klasifikasi yang benar tidak memiliki harga dan hasil klasifikasi yang tidak benar memiliki harga yang linier. Demikian pula, C(0,n^2), C(0,2^n) and C(0,n^n) merepresentasikan harga harga dimana klasifikasi yang benar tidak memiliki harga, sementara hasil klasifikasi yang tidak benar memiliki harga yang kuadratik dan eksponensial.

Dengan menggunakan harga matriks C(0,n^n), kita dapat melihat bahwa model yang penulis kembangkan hampir selalu mengungguli model selective Naive Bayes dalam tingkat akurasi dan biaya. Meskipun model model ini mendapatkan akurasi paling tinggi(0.504) dalam tahun pertama , algoritma baru yang penulis kembangkan , khususnya CS-SNB-Accuracy, mendapatkan tingkat akurasi paling tinggi pada tahun kedua (0.518), tahun ketiga (0.542) dan tahun keempat (0.532). Dengan biaya rata-rata, kita dapat melihat bahwa model yang dikembangkan penulis khususnya CS-SNB-Cost, selalu mendapatkan biaya yang lebih rendah dibandingkan selective Naive-Bayes. [40][42][43][44][45][46]

Berfokus pada harga matriks, kita dapat melihat bahwa akurasi bervariasi dari setiap model dan tahun prediksi, tetapi penulis tidak melihat adanya pola. Model selective naive Bayes mendapatkan akurasi paling tinggi di tahun pertama(0.507) dengan menggunakan biaya matriks C(0,2^n). Sementara model CS-SNB-Accuracy mendapatkan tingkat akurasi paling tinggi di tahun kedua dan tiga dengan menggunakan harga matriks C(0,2^n) dan C(0,n^n). Dan model CS-SNB-Cost mendapatkan tingkat akurasi paling tinggi pada tahun keempat dengan menggunakan biaya matriksC(0,n^2). Dari biaya biaya, kita dapat mengetahui bahwa buaya rata rata terendah dan tertinggi di dapatkan oleh C(0,n) dan C(0,2^n). CS-SNB-Cost hampir mendapatkan harga paling rendah dengan menggunakan matriks C(0,n) .

Perihal setiap algoritma, kita dapat mengetahui bahwa Naive Bayes mendapatkan akurasi paling tinggi untuk model pada tahun pertama dengan menggunakan biaya matriks apapun. Sementara CS-SNB-Accuracy dan CS-SNB-Cost memprediksi hampir semua nilai lebih akurat dibandingkan model Naive Bayes pada tahun tahun berikutnya.

Table 3 menunjukkan bahwa akurasi dan harga rata-rata dari berbagai model klasifikasi yang dilatih menggunakan biaya matriks C(0,n^n) untuk tahun tahun prediksi.

Dengan menganalisa tingkat keakuratan, kita dapat membagi menjadi tiga kelompok yang berbeda (tingkat rendah, sedang , dan tinggi). Kelompok pertama terdapat klasifikasi Naive Bayes yang mendapatkan nilai yang rendah, Untuk di kelompok kedua, terdiri 3 klasifikasi (selective naive Bayes, cost-sensitive naive Bayes-Accuracy, dan cost-sensitive naive Bayes-Cost) yang mendapatkan nilai pada tingkat medium, sementara kelompok ketiga terdiri dari klasifikasi non-Bayesian (C4.5 decision tree, K-nearest neighbour and logistic regression) yang memiliki nilai tertinggi. Kita dapat melihat diatas bahwa sifatnya sama tidak peduli tahun prediksi yang digunakan. [47][48][49] [50] Hasil dari tes Kruskal-Wallis menunjukkan bahwa ada perbedaan yang signifikan antara 7 klasifikasi yang berbasis akurasi. Untuk itu penulis menggunakan tes Mann-Whitney untuk mencari tahu klasifikasi mana yang memiliki tingkat yang lebih baik berdasarkan criterianya. Penulis membandingkan klasifikasi yang memilikitingkat rata rata yang paling tinggi dibandingkan yang lain. Klasifikasi ini ditandai pada tabel 3 dengan menggunakan simbol † .

Mengenai nilai dari biaya, ktia juga membedakan berdasarkan 3 kelompok, dalam hal ini , naive-Bayes dan selective naive-Bayes mendapatkan biaya yang lebih tinggi, dimana C4.5 decision tree, K-nearest neighbour dan logistic regression mendapatkan biaya dengan tingkat medium. Dan yang terakhir CS-SNB-Accuracy dan CS-SNB-Cost mendapatkan biaya yang paling rendah. Kita juga mencoba tes Kruskal-Wallis utnuk membandingkan tiap tiap klasifikasi yang berbasis biaya rata rata.

Dengan menganalisa klasifikasi yang menggunakan cost-sensitive, kita dapat membandingkan algoritma yang penulis rancang dengan algoritma MetaCost, CostSensitiveClassifier , dan CSRoulette. Klasifikasi ini melakukan konversi terhadap klasifikasi yang cost-insensitive menjadi cost-sensitive. Tabel 4 menunjukkan akurasi dan biaya rata-rata dari klasifikasi diatas yang dilatih menggunakan biaya matriks C(0,2^n) untuk semua tahun prediksi. [51][52][53]

Table 2
Accuracy and average cost of models which are learned using different selective naive Bayes approaches and cost matrices.

Methods	First year		Second year		Third year		Fourth year	
	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
Cost matrix: $C(0,n)$								
Selective naive Bayes	**0.502**	0.608	**0.506**	0.664	0.530	0.563	0.517	0.584
CS-SNB-Accuracy	0.477	0.610	0.513	0.579	0.530	0.543	0.528	**0.534**
CS-SNB-Cost	0.458	**0.586**	0.519	**0.577**	**0.534**	**0.538**	**0.532**	0.548
Cost matrix: $C(0,n^2)$								
Selective naive Bayes	**0.503**	0.828	0.501	1.005	0.525	0.758	0.532	0.742
CS-SNB-Accuracy	0.460	**0.721**	0.498	0.873	0.565	0.766	0.525	0.713
CS-SNB-Cost	0.451	0.735	**0.504**	**0.775**	**0.532**	**0.708**	**0.533**	**0.706**
Cost matrix: $C(0,2^n)$								
Selective naive Bayes	**0.507**	1.211	0.509	1.299	0.504	1.171	0.508	1.170
CS-SNB-Accuracy	0.480	1.221	**0.519**	**1.156**	**0.538**	**1.069**	0.523	1.101
CS-SNB-Cost	0.460	**1.187**	0.507	1.168	0.533	1.080	**0.532**	**1.086**
Cost matrix: $C(0,n^n)$								
Selective naive Bayes	**0.504**	1.227	0.506	1.327	0.526	1.133	0.516	1.190
CS-SNB-Accuracy	0.446	0.953	**0.518**	**0.769**	**0.542**	0.714	**0.532**	0.732
CS-SNB-Cost	0.439	**0.753**	0.500	0.772	0.513	**0.695**	0.516	**0.705**

Table 3
Accuracy and average cost of models which are learned using different classification methods. Results are achieved using the cost matrix $C(0,n^n)$ for all prediction years.

Methods	First year		Second year		Third year		Fourth year	
	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
NB	0.262†	3.138†	0.343†	2.815†	0.306†	2.719†	0.303†	2.779†
SNB	0.504†	1.227†	0.506†	1.327†	0.526†	1.133†	0.516†	1.190†
CS-SNB-Accuracy	0.446†	0.953†	0.518†	**0.769**	0.542†	0.714	0.532†	0.732
CS-SNB-Cost	0.419†	**0.753**	0.500†	0.772	0.513†	**0.695**	0.516†	**0.705**
C4.5	0.525†	1.204†	0.598	1.073†	0.640	0.793†	0.634	0.879†
K-NN	0.553†	1.113†	**0.609**	1.080†	**0.643**	0.803†	**0.655**	0.857†
Logistic	**0.587**	0.978†	0.587†	1.058†	0.622	0.866†	0.625	0.878†

Naive Bayes (NB); Selective naive Bayes (SNB); C4.5 decision tree (C4.5); K-nearest neighbour (K-NN); Logistic regression (Logistic).

Table 4
Accuracy and average cost of models which are learned using different cost-sensitive approaches. Values achieved using the cost matrix $C(0,2^n)$ for all prediction years.

Methods	First year		Second year		Third year		Fourth year	
	Accur	Cost	Accur	Cost	Accur	Cost	Accur	Cost
MetaCost	0.116†	1.770†	0.315†	1.597†	0.326†	1.347†	0.188†	1.632†
CostSensitiveClassifier	0.253†	1.866†	0.329†	2.382†	0.300†	2.005†	0.238†	1.843†
CSRoulette	0.432†	2.878†	0.517	2.923†	0.521†	2.623†	0.526	2.893†
CS-SNB-Accuracy	**0.480**	1.221†	**0.519**	**1.156**	**0.538**	**1.069**	0.523	1.101
CS-SNB-Cost	0.460†	**1.187**	0.507†	1.168	0.533	1.080	**0.532**	**1.086**

Table 5
Variables, accuracy and cost for the *CS-SNB-Cost* model by each fold of the cross-validation process. Values achieved using the cost matrix $C(0,n^2)$ for all prediction years.

k	First year			*k*	Second year		
	Variables	Accuracy	Cost		Variables	Accuracy	Cost
1	12,11	0.456	0.721	1	12,11,14	0.497	0.866
2	12,11	0.478	0.756	2	12,16,14	0.566	0.783
3	12,11	0.465	0.713	3	12,16,14	0.537	0.733
4	12,14	0.391	0.834	4	12,16,14	0.492	0.847
5	12,14	0.521	0.647	5	12,16	0.507	0.783
6	12,11	0.456	0.713	6	12,11,14	0.492	0.788
7	12,11	0.439	0.730	7	12,16,14	0.537	0.793
8	12,11	0.447	0.682	8	12,16,14	0.566	0.596
9	12,14	0.428	0.765	9	12,16	0.522	0.778
10	12,11	0.426	0.782	10	12,16,14	0.458	0.778
Mean values		**0.451**	**0.735**			**0.514**	**0.775**
k	Third year			*k*	Fourth year		
	Variables	Accuracy	Cost		Variables	Accuracy	Cost
1	12,14,16	0.522	0.707	1	12,14,16	0.551	0.662
2	12,11,14	0.500	0.797	2	12,14,6	0.564	0.759
3	12,11,14	0.544	0.623	3	16,12,14	0.493	0.753
4	12,16,14	0.533	0.752	4	12,16,14	0.525	0.701
5	12,14,7	0.533	0.685	5	12,16,14	0.558	0.636
6	12,14,6	0.561	0.725	6	12,16,14	0.538	0.655
7	12,11,14	0.556	0.646	7	12,16,14	0.545	0.668
8	12,14,7	0.522	0.752	8	12,14,4	0.506	0.766
9	12,14,4	0.511	0.707	9	12,16,14	0.493	0.798
10	12,14,4	0.533	0.634	10	12,14,6	0.538	0.655
Mean values		**0.532**	**0.708**			**0.533**	**0.706**

Berfokus pada akurasi dan biaya, kita dapat melihat bahwa pada tabel 4 , model yang dirancang oleh penulis lebih unggul dibandingkan klasifikasi yang lain. Hasil dari tes Kruskal-Waillis menunjukkan bahwa ada perbedaan signifikan antara klasifikasi yang berbasis akurasi dan biaya. Untuk itu penulis menggunakan tes Mann-Whitney untuk mencari tahu klasifikasi mana yang lebih baik berdasarkan

kriteria ini. Penulis membandingkan klasifikasi yang memiliki nilai rata-rata terbaik. Klasifikasi yang bertanda pada table 4 dengan simbol † memiliki perbedaan yang signifikan berdasarkan tingkat rata-rata. Hasil juga menunjukkan bahwa ada perbedaan yang signifikan antara CS-SNB-Accuracy dan MetaCost, CostSensiticeClassifier, CSRoulette dan CS-SNB-Cost dalam hal akurasi. Hasil juga menunjukkan bahwa ada perbedaan yang signifikan antara CS-SNB-Cost dan MetaCost, CostSensiticeClassifier, CSRoulette dan CS-SNB-Accuracy dalam hal biaya.

Untuk itu kita dapat menganalisis model lebih detail. Pada tabel 5 menunjukkan variabel, akurasi , dan biaya yang spesifik untuk model CS-SNB-Cost yang telah melewati hasil validasi silang. Nilai-nilai ini didapatkan dengan menggunakan biaya matriks $C(0,n^2)$. Dengan menganalisis tabel 5, kita dapat menemukan bahwa model-model selalu terdapat dalam impact factor. Model-model biasanya juga biasanya mengikutsertakan variabel seperti h_c-indeks , half life dan the article influence. Kita juga dapat mencatat bahwa model pada tahun pertama selalu memiliki 2 variabel sementara tahun kedua, ketiga dan keempat selalu memiliki 3 variabel.

KESIMPULAN

Jurnal ini menyajikan algoritma baru untuk memprediksi peningkatan indeks-h untuk jurnal ilmiah berdasarkan pendekatan cost-sensitive dan pemilihan subset fitur. Kami mengembangkan metode cost-sensitive yang berbeda, di mana algoritma pembelajaran mencakup biaya misklasifikasi. Pendekatan ini memperhitungkan biaya misklasifikasi berbeda dari 0 (hit) dan 1 (miss). Algoritma ini berkaitan dengan akurasi klasifikasi dan biaya klasifikasi. Khususnya, penulis mengembangkan dua pendekatan cost-sensitive selective naive Bayes. Proses pencarian dari pendekatan pertama (CS-SNB-Accuracy) termasuk variabel yang meningkatkan akurasi klasifikasi, sedangkan proses pencarian dari pendekatan kedua (CS-SNB-Cost) termasuk variabel yang mengurangi jarak antara kelas yang sebenarnya dan yang diprediksi. Tujuan utama dari algoritma yang diusulkan adalah untuk memprediksi peningkatan tahunan indeks-h untuk jurnal ilmiah. Model-model tersebut mampu memprediksi indeks-h dimana jurnal ilmiah dapat menjadi alat yang berguna untuk komunitas ilmiah.

Hasil menunjukkan bahwa pendekatan penulis, khususnya Akurasi CS-SNB, mencapai nilai akurasi yang lebih tinggi daripada klasifikasi cost-sensitive dan klasifikasi Bayesian. Selain itu, kami juga mencatat bahwa CS-SNB-Cost menghasilkan rata rata biaya yang lebih rendah dari semua kerangka pengukuran. Algoritma Cost-sensitive selective naive Bayes ini mengungguli algoritma selective naive Bayes yang asli dalam hal akurasi dan biaya rata-rata, sehingga pendekatan pembelajaran cost-sensitive dapat digunakan dalam berbagai pendekatan klasifikasi probabilistik.

Di masa depan, penulis bertujuan untuk membangun klasifikasi cost-sensitive Bayesian baru seperti selective tree augmented naive Bayes. Sehingga pada akhirnya, nilai indeks-h bisa bervariasi tergantung pada sumber yang dikonsultasikan (Google Scholar, Scopus, ISI WoK, dll.).

Daftar Pustaka

[1] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, h-index: a review focused in its variants, computation and standardization for different scientific fields, J. Informetr. 3 (4) (2009) 273–289.

[2] S. Alonso, F.J. Cabrerizo, E. Herrera-Viedma, F. Herrera, hg-index: a new index to characterize the scientific output of researchers based on the h- and g-indices, Scientometrics 82 (2) (2010) 391–400.

[3] O.K. Baskurt, Time series analysis of publication counts of a university: what are the implications? Scientometrics 86 (3) (2011) 645–656.

[4] P.D. Batista, M.G. Campiteli, O. Kinouchi, A.S. Martinez, Is it possible to compare researchers with different scientific interests? Scientometrics 68 (1) (2006) 179–189.

[5] L. Bornmann, R. Mutz, H. Daniel, Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine, J. Am. Soc. Inf. Sci. Technol. 59 (5) (2008) 830–837.

[6] F.J. Cabrerizo, S. Alonso, E. Herrera-Viedma, F. Herrera, q2-index: quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core, J. Informetr. 4 (1) (2010) 23–28.
[7] J.S. Cardodo, J.F.P. da Costa, Learning to classify ordinal data: the data replication method, J. Mach. Learn. Res. 8 (2007) 1393–1429.
[8] K. Crammer, Y. Singer, Pranking with ranking, in: Advances in Neural Information Processing Systems, vol. 14, 2002, MIT Press, pp. 641–647.
[9] P. Domingos, Metacost: a general method for making classifiers cost-sensitive, in: Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, 1999, pp. 155–164.
[10] C. Drummond, R. Holte, Exploiting the cost (in)sensitivity of decision tree splitting criteria, in: Proceedings of the 17th International Conference on Machine Learning, 2000, pp. 239–246.
[11] D.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley,New York, USA, 1973.
[12] L. Egghe, Dynamic h-index: the Hirsch index in function of time, J. Am. Soc. Inf. Sci. Technol. 58 (3) (2006) 452–454.
[13] L. Egghe, An improvement of the h-index: The g-index, ISSI Newslett. 2 (1) (2006) 8–9.
[14] L. Egghe, The hirsch-index are related impact measures, Annu. Rev. Inf. Sci. Technol. 44 (2010) 65–114.
[15] L. Egghe, R. Rousseau, An informetric model for the hirsch-index, Scientometrics 69 (1) (2006) 121–129.
[16] C. Elkan, The foundations of cost-sensitive learning, in: Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence, 2001, pp. 973–978.
[17] E. Frank, M. Hall, A simple approach to ordinal classification, in: Proceedings of the 12th European Conference on Machine Learning, 2001, pp. 145–156.
[18] E. Frank, S. Kramer, Ensembles pf nested dichotomies for multi-class problems, in: Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 305–312.
[19] J. Furnkranz, Pairwise classification as an ensemble technique, in: Proceedings of the 13th European Conference on Machine Learning, 2002, pp. 97–110.
[20] P.E. Hart, The condensed nearest neighbour rule, Trans. Inf. Theory 14 (1968) 515–516.
[21] R. Herbrich, T. Graepel, K. Obermayer, Regression Models for Ordinal Data: A Machine Learning Approach. Technical Report 99-3, Department of Computer Science, Technical University of Berlin, 1999.
[22] R. Herbrich, T. Graepel, K. Obermayer, Large margin rank boundaries for ordinal regression, in: Advances in Large Margin Classifiers, MIT Press, Cambridge, MA, 2000, pp. 115–132 (Chapter 7).
[23] F. F. Tampinongkol, Y. Herdiyeni, and E. N. Herliyana, "Feature extraction of Jabon (Anthocephalus sp) leaf disease using discrete wavelet transform," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, p. 740, Apr. 2020, doi: 10.12928/telkomnika.v18i2.10714.
[24] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, 2nd ed., Wiley, New York, USA, 2000.
[25] A. Ibáñez, P. Larrañaga, C. Bielza, Predicting citation count of bioinformatics papers within four years of publication, Bioinformatics 25 (24) (2009) 3303–3309.
[26] K. L. Hartono and J. A. Ginting, "Implementation of web-based Japanese digital handwriting OCR using chain code and manhattan distance," 2023, p. 020017. doi: 10.1063/5.0174709.
[27] B. Jin, h-index: an evaluation indicator proposed by scientist, Sci. Focus 1 (1) (2006) 8–9.
[28] S.B. Kotsiantis, Local ordinal classification, in: Artificial Intelligence Applications and Innovations. International Federation for Information Processing, Springer, Athens, Greece, 2004, pp. 1–8.
[29] S.B. Kotsiantis, P.E. Pintelas, A cost sensitive technique for ordinal classification problems, in: Methods and Applications of Artificial Intelligence. Lecture Notes in Computer Science, Springer, Samos, Greece, 2004, pp. 220–229.

[30] C. Herdian, S. Widianto, J. A. Ginting, Y. M. Geasela, and J. Sutrisno, "The Use of Feature Engineering and Hyperparameter Tuning for Machine Learning Accuracy Optimization: A Case Study on Heart Disease Prediction," 2024, pp. 193–218. doi: 10.1007/978-3-031-50300-9_11.
[31] G. Krampen, A. von Eye, G. Schui, Forecasting trends of development of psychology from a bibliometric perspective, Scientometrics 87 (2) (2011) 687–694.
[32] W. Kruskal, W. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (260) (1952) 583–621.
[33] P. Langley, S. Sage, Induction of selective bayesian classifiers, in: Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence, 1994, pp. 399–406.
[34] H.T. Lin, L. Li, Reduction from cost-sensitive ordinal ranking to weighted binary classification, Neural Comput. 24 (5) (2012) 1329–1367.
[35] C.X. Ling, Q. Yang, J. Wang, S. Zhang, Decision trees with minimal costs, in: Proceedings of the 21st International Conference on Machine Learning, 2004, pp. 69–77.
[36] H. Mann, D. Whitney, On a test of whether one of two random variables is stochastically larger than the other, Ann. Math. Stat. 18 (1) (1947) 50–60.
[37] P. McCullagh, Regression models for ordinal data, J. R. Stat. Soc. Ser. B 42 (2) (1980) 109–142.
[38] P. McCullagh, J.A. Nelder, Generalized Linear Models, Chapman and Hall, London, 1983.
[39] M. Minsky, Steps toward artificial intelligence, IRE 49 (1) (1961) 8–30.
[40] R. Potharst, J.C. Bioch, Decision trees for ordinal classification, Intell. Data Anal. 4 (2) (2000) 97–112.
[41] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco, USA, 1993.
[42] F. Ruane, R.S.J. Tol, Rational (successive) h-indices: an application toeconomics in the Republic of Ireland, Scientometrics 75 (2) (2008) 395–405.
[43] A. Shashua, A. Levin, Ranking with large margin principle: two approaches, in: Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, MA, 2003, pp. 961–968.
[44] V.S. Sheng, C.X. Ling, Roulette sampling for cost-sensitive learning, in: Proceedings of the 18th European Conference on Machine Learning. Lecture Notes in Computer Science, 2007, Springer, pp. 724–731.
[45] A. Sidiropoulos, D. Katsaros, Y. Manolopoulos, Generalized hirsch h-index for disclosing latent facts in citation networks, Scientometrics 72 (2) (2007) 253–280.
[46] M. Stone, Cross-validation choice and assessment of statistical predictions, J. R. Stat. Soc. 36 (1974) 111–147.
[47] K.M. Ting, Inducing cost-sensitive trees via instances weighting, in: Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, 1998, pp. 23–26.
[48] P.D. Turney, Cost-sensitive classification: empirical evaluation of a hybrid genetic decision tree induction algorithm, J. Artif. Intell. Res. 2 (1995) 369–409.
[49] I.H. Witten, E. Frank, Data Mining—Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann Publishers, San Francisco, CA, 2005.
[50] F.Y. Ye, R. Rousseau, The power law model and total career h-index sequences, J. Informetr. 2 (4) (2008) 288–297.
[51] B. Zadrozny, C. Elkan, Learning and making decisions when costs and probabilities are both unknown, in: Proceedings of the 7th International Conference on Knowledge Discovery and Data Mining, 2001, pp. 204–213.
[52] B. Zadrozny, J. Langford, N. Abe, Cost-sensitive learning by cost-proportionate instance weighting, in: Proceedings of the 3rd International Conference on Data Mining, 2003, pp. 435–442.